Award Number: DAMD17-01-1-0376

TITLE: Investigating the Mechanisms of Action and the Identification of Breast Carcinogens by Computational Analysis of Female Rodent Carcinogens

PRINCIPAL INVESTIGATOR: Albert R. Cunningham, Ph.D.

CONTRACTING ORGANIZATION: Louisiana State University
Baton Rouge, Louisiana 70803-2701

REPORT DATE: August 2004

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 2004	Annual (15 Jul 2003 - 14 Jul 2004)

4. TITLE AND SUBTITLE
Investigating the Mechanisms of Action and the Identification of Breast Carcinogens by Computational Analysis of Female Rodent Carcinogens

5. FUNDING NUMBERS
DAMD17-01-1-0376

6. AUTHOR(S)
Albert R. Cunningham, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Louisiana State University
Baton Rouge, Louisiana 70803-2701

E-Mail: arc@lsu.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

This project is investigating the potential that environmental estrogens may be involved in the etiology of breast cancer. We hypothesize that specific features of chemicals can be identified that are significantly associated with female and breast carcinogens and that these features are related to mechanisms of chemical carcinogenesis. Our overall scientific objective is to investigate the hypothesized relationship between environmental chemicals, xenoestrogens, and the development of breast cancer.

With the success of the rat and mouse mammary carcinogen models we are preparing two manuscripts for publication. We are also pursuing work on a general chemical carcinogen manuscript and a one describing female-specific carcinogens. Also of importance, we are working on several xenoestrogen models that, although not detailed in the project proposal, will be of great importance for understanding the endocrine disruptor link to breast cancer. We have also developed a new structure-activity relationship program called cat-SAR that is producing predictive and mechanistically insightful models of mammary carcinogens. Looking forward I see no obstacles to the successful completion of this project in a timely manner.

14. SUBJECT TERMS
Structure-activity relationship (SAR), computer modeling, mechanisms and etiology of breast cancer, environmental carcinogens

15. NUMBER OF PAGES
57

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Annual Review August 2004

Investigating the Mechanism of Action and the Identification of Breast Carcinogens by Computational Analysis of Female Rodent Carcinogens

DAMD17-01-1-0376

Albert R. Cunningham, Ph.D.

Introduction

The well-established breast cancer risk factors may account for only 47% of the breast cancer incidence in the United States. This leaves a considerable portion of breast cancer from undetermined origin. This project is investigating the potential that environmental chemicals and particularly those with estrogenic activity may be involved in the etiology of breast cancer. We hypothesize that specific features of chemicals can be identified that are significantly associated with female and breast carcinogens and that these features are related to mechanisms of chemical carcinogenesis. Our overall scientific objective is to investigate the hypothesized relationship between environmental chemicals, xenoestrogens, and the developmental of breast cancer. The successful completion of this project will provide mechanistic information related to chemical-induced breast cancer as well as structure-activity relationship (SAR) models capable of estimating the likelihood that chemicals with unknown carcinogenic activity may be breast carcinogens.

Body

Software Change

SAR modeling for this project was originally proposed to be conducted with the MCASE program. However, for multiple reasons I have decided to develop our own system. This change does not alter the project and I am currently working with LSU's Sponsored Program Administrator to update the Statement of Work. I have discussed this matter with Dr. Moore. I am including the updated (but not yet approved) Statement of Work in the appendices.

During the early part of the project it was becoming evident that MCASE was not developing models that were of stellar predictivity. On account of successful modeling with SIMCA (soft independent modeling of class analogy) of aromatic amine *Salmonella* mutagens and skin sensitizing agents for a project supported by Proctor & Gamble, we spent some time investigating whether SIMCA models could be employed to produce adequate models relating to this project.

We originally thought that SIMCA combined with HQSAR (hologram quantitative SAR) models appeared to be superior to MCASE models. The HQSAR-SIMCA approach utilized categorical biological data (i.e., carcinogen vs. non-carcinogen) and molecular fragments as SAR descriptors. Therefore, this seemed a reasonable substitute SAR approach for MCASE. However, upon consultation with the makers of Sybyl HQSAR-SIMCA, we learned that there was a large degree of random assignment of SAR descriptors. Basically, as it turned out, although the modeling software was able to produce models that could predict the activity of unknown chemicals—they were not very

mechanistically insightful. In other words, we would not be able to interpret these models in order to understand the structural attributes of breast carcinogens. Moreover, without an SAR model having a solid and understandable mechanistic foundation, we were troubled that the even the "predictive" models may have more to do with chance occurrences than true accurate predictions.

At this point I was concerned about the completion of the project. However, we have been successful in developing a new SAR system that we are calling cat-SAR for categorical SAR. I discussed the development of this new program with Dr. Moore—and we were in agreement that this would be an appropriate path to take in order to achieve the overall goals of the project. The program has been developed with guidance from Prof. Herb Rosenkranz. Dr. Rosenkranz is co-PI of this project and a co-developer of MCASE.

I have one publication that has just been returned to the editor of *SAR and QSAR in Environmental Research* along with the requested revisions. This manuscript described the program in detail. We note the publication was on respiratory sensitizers—not breast carcinogens. The reason for this was 1) it was a small and manageable dataset and 2) a previous MCASE analysis of this data yielded a very good model. As such, this was a suitable dataset on which to develop and test the cat-SAR program. A copy of the manuscript detailing the cat-SAR program is included in the appendices.

Specific Aim Accomplishments
The Specific Aims for year one are as follows:

Specific aim 1: Development and validation of SAR models for female breast carcinogens (months 1-12).

a. Identify chemicals tested in female rodents from the Carcinogenic Potency Database and the National Toxicology Program (month 1).
b. Enter chemical structures and potency values into MCASE program (months 2-8).
c. Validate models using 10-fold cross validation (months 9-12).
d. Summarize and interpret models and prepare publication.

These models have been developed and validated (i.e., a-c) as planned in MCASE as previously reported. Within the last year, they have now been developed with the new cat-SAR program. We have also updated rodent carcinogenicity models so that all models (mouse and rat, as well and female specific version) have been built on the same datasets and analyzed with the cat-SAR program. We are preparing to publish two manuscripts describing mouse and rat mammary carcinogens.

Female Carcinogen Models
Specific Aim 1a is for the creation of female specific models. As discussed, these models have been developed for MCASE. We are preparing to import these models to cat-SAR for analysis.

Mammary Carcinogen Models

We have had great success in developing the mammary carcinogen model. So much so, in fact, we started questioning the validity of the models. With respect to this, we have devoted a significant effort in "assuring" their appropriateness. Basically, we have now developed several different mammary carcinogen models. One set is based on mouse mammary carcinogens and one set is based on rat.

1. Rat Mammary Carcinogen Models: From the published CPDB target site summary (*15*) we developed a SAR learning set of 100 compounds shown to induce breast cancer in rat. The cat-SAR program develops SAR models through the comparison of structural features associated with categorical responses (e.g., active and inactive compounds). When just considering carcinogenesis, the categories are clearly carcinogens and noncarcinogens. However, when considering organ-specific carcinogenesis, the question arises as to the selection of the inactive compounds. Should they be noncarcinogens or carcinogens that are just not carcinogenic to the organ under study? For this exercise we considered both options. (We note this important aspect of the project was not considered in the original proposal.) Thus we developed two separate models for rat mammary carcinogens: The mammary carcinogen - noncarcinogen model and the mammary carcinogen - non-mammary carcinogen model.

Additionally, since the CPDB lists 449 compounds as rat noncarcinogens we had a choice of noncarcinogens to include on the model. Likewise, the CPDB also lists 495 carcinogens, 395 of which do not induce cancer in the mammary gland. For each model we randomly sampled the inactive datasets to derive three sets (designated Models 1, 2 and 3 in Tables 1 and 2) of 100 chemicals each in which to balance the 100 rat breast carcinogens. We did this to assure that our models were in fact accurately describing mammary carcinogens—not just chance occurrences. (We note again this important aspect of the project was not considered in the original proposal.) Statistical comparison of the each of the model's fragment sets and predictivity was conducted to determine whether the three sets were statistically different.

Tables 1 and 2 (at the end of this report) are from a manuscript being prepared detailing the rat mammary carcinogen models. Of particular note is that each set of models demonstrates predictivity for unknown compounds in the 70-80% accuracy range (observed correct prediction rate or OCP). We are currently in the process of a mechanistic analysis of the models in order to identify and understand molecular attributes of breast carcinogens.

2. Mouse Mammary Carcinogen Models:

A similar group of analyses as listed above for rat carcinogens is being completed for mouse carcinogens. The validation results are shown in Tables 3 and 4. These models are based on 24 mouse mammary carcinogens from the published CPDB target site summary (*15*).

The Second Specific Aim is:

Specific aim 2: Identify chemical and biological attributes of female and/or breast carcinogens to provide evidence to test the hypothesis that xenoestrogens are involved in breast cancer (months 13-36).

a. Compare and identify Structural Feature Overlap Method of female and breast carcinogens to those of other available toxicological SAR models (see Facilities and Equipment for a complete list of available models) (months 13-16).
b. As above using Joint Prevalence Method (months 16-24).
c. Identify the exact features of female and breast carcinogen models that are responsible for predicted similar activities identified above (months 25-26).
d. Conduct QSAR and CoMFA analyses with chemicals containing these structures using biological data from appropriate assays (months 28-36).
e. Conduct metabolism experiments on identified outliers to see whether metabolic activation is required for activity and update models if required (months 28-36).
f. Summarize and interpret data and prepare publications (months 28-36).

We are just concluding migration of a set of about 20 MCASE toxicological SAR models to cat-SAR. We are in the process of validating these models. This is required for Specific Aim 1a and will be done shortly. Of particular interest is the fact that we are developing three estrogen cat-SAR models that will be directly applicable to testing the relationship between estrogenicity and mammary carcinogenicity.

We have also nearly completed the development of a new method, applicable to the cat-SAR program for comparing joint prevalence (i.e, toxicological mechanism similarity) of SA 2b. We note we can still perform this analysis with the previously published method since it only requires estimations of toxicity—which do not require MCASE but can use cat-SAR derived values.

SA 2c,d, and f should be accomplished successfully in the upcoming year. We note that SA 3e requires the MCASE module META. We do not have current access to a working copy of MCASE, though Professor Rosenkranz is working on getting a copy for use with this project.

Key Research Accomplishments

Developed new SAR modeling algorithm called cat-SAR.

Developed predictive and mechanistically insightful SAR models for rat and mouse carcinogens and mammary carcinogens

Development of shareable databases/learning sets of chemical carcinogens, their molecular structure and associated activity values

Reportable Outcomes

Seminars

"Structure-activity relationships: Estrogen mimics and endocrine disruptors" presented to Professor John McLachlan's research group at Tulane University

Manuscripts

Cunningham AR, Cunningham SL, Consoer DM, Moss ST, Karol MH. Development of an information-intensive structure- activity relationship model and its application to human respiratory chemical sensitizers. SAR and QSAR in Environmental Research (pending acceptance of revisions)

Cunningham AR, Cunningham SL, Rosenkranz HR. Structure Activity Approach to the Identification of Environmental Estrogens: The MCASE Approach. SAR and QSAR in Environmental Research 15:55-67(2004).

Patent/copyright

We have submitted a patent and copyright application to LSU's Office of Intellectual Property for the cat-SAR computational toxicology expert system. As noted, this system was developed to replace the MCASE system described in the original proposal and SOW.

Funding Applied for Based on Work Supported by this Award

We note that the below listed proposals all relate to the discovery of novel antibreast cancer therapeutics. Given that the estrogen receptor is involved in the etiology, cure, and prevention of breast cancer, this IDEA Award has allowed us to pursue new avenues of research into drug discovery.

AWARDED

Pharmacophore discovery by differential toxicity studies, LSU Faculty Research Grant, (PI, $10,000)

PENDING

A novel approach for the identification of pharmacophores through differential toxicity analysis of estrogen receptor positive and negative cell lines, Department of Defense Breast Cancer Research Program, pending (PI, $372,542)

Pharmacophore discovery by differential toxicity studies, National Institutes of Health, pending (PI, $1,052,735)

NOT AWARDED

Identification of pharmacophores through differential toxicity analysis: American Cancer Society, submitted, 2003 (PI, $748,010)

Comment summary: The proposal was ranked 6th and not funded.

Identification of pharmacophores through differential toxicity analysis: Louisiana Board of Regents Research and Development Program Research Competitiveness Subprogram, submitted, 2003 (PI, $180,000 with $65,000 LSU match)

TO BE SUBMITTED

We note, beginning prior to July 14, 2004, we have started the preparation of two additional proposals that are directly related to the successful development of mammary carcinogen models. These proposals, which will be described in detail for the next report, employ the mammary carcinogen models with geographic information system (GIS) analysis. These projects will look at breast cancer rates related to toxic release inventory (TRI) data. We will use the models developed in this IDEA award to estimate the breast cancer potential of TRI chemicals.

Conclusions

With the success of the rat and mouse mammary carcinogen models we are preparing two manuscripts for publication. We are also pursuing work on a general chemical carcinogen manuscript and a one describing female-specific carcinogens. Also of importance, we are working on several xenoestrogen models that, although not detailed in the project proposal, will be of great importance for understanding the endocrine disruptor link to breast cancer.

To date, after technically about two years of work we have developed the proposed models set forth in Specific Aim 1 using MCASE. We are now slightly behind schedule due to the time required to develop the cat-SAR program. However, in conjunction with this and other projects in my laboratory, all the required components for Specific Aim 2 are being moved from MCASE to cat-SAR. There should be no significant future delays or problems accomplishing the tasks of Specific Aim 2. This is of particular relevance for Specific Aims 2a and 2b that require other toxicological models (e.g., mutagenicity and estrogenicity) on which to compare the female and mammary gland carcinogen models.

Looking forward I see no obstacles to the successful completion of this project in a timely manner. However, we may request a no cost extension. With the transition to cat-SAR we envision being able to more accurately and thoroughly investigate the chemical structural attributes of breast carcinogens as well as produce the required predictive models for estimating the mammary cancer causing potential of untested chemicals.

TABLE 1. Predictive performance summary for rat mammary carcinogen – nonmammary carcinogen SAR model. The ABC model was based on fragments of size between three and seven heavy atoms and considered atoms, bonds, and atom connection. The ABCH model also included consideration of hydrogen atoms.

Model	*Total. Fragments*	*Model Fragments*	*Active Fragments*	*Inactive Fragments*	*Sensitivity*	*Specificity*	*OCP#*
ABC3/0.75							
Model 1	13868	1349	849	500	0.80(70/88)	0.66(53/80)	0.73(123/168)
Model 2	14461				0.72(63/87)	0.72(59/82)	0.72(122/169)
Model 3	14427	1245	767	478	0.68(59/87)	0.74(64/86)	0.71(123/173)
ABC3/0.90							
Model 1	13868	1102	731	371	0.83(58/70)	0.74(40/54)	0.79(98/124)
Model 2	14461				0.82(54/66)	0.72(44/64)	0.77(98/130)
Model 3	14427	847	520	327	0.82(51/62)	0.72(41/57)	0.77(92/119)
ABCH3/0.75							
Model 1	32235	3679	2081	1598	0.81(78/96)	0.62(55/89)	0.72(133/185)
Model 2	32374	3921	2088	1833	0.70(66/94)	0.64(59/92)	0.67(125/186)
Model 3	32627	3497	1928	1569	0.75(70/93)	0.69(65/94)	0.72(135/187)
ABCH3/0.90							
Model 1	32235	2750	1642	1108	0.81(65/80)	0.76(50/66)	0.79(115/146)
Model 2	32374	2947	1637	1310	0.75(55/73)	0.69(53/77)	0.72(108/150)
Model 3	32627	2241	1170	1071	0.81(63/78)	0.70(52/74)	0.76(115/152)

Footnotes:
Total Fragments: fragments derived from learning set.
Model Fragments: fragments meeting specified rules of the model.
Active Fragments: fragments meeting specified rules to be considered as active.
Inactive Fragments: fragments meeting specified rules to be considered as inactive.
Sensitivity: number of correct positive predictions / total number of positives.
Specificity: number of correct negative predictions / total number of negatives.
OCP: number of correct predictions / total number of predictions.

TABLE 2. Predictive performance summary for rat mammary carcinogen – noncarcinogen SAR model. The ABC model was based on fragments of size between three and seven heavy atoms and considered atoms, bonds, and atom connection. The ABCH model also included consideration of hydrogen atoms.

Model	*Total. Fragments*	*Model Fragments*	*Active Fragments*	*Inactive Fragments*	*Sensitivity*	*Specificity*	*OCP#*
ABC3/0.75							
Model 1	18021	1336	758	578	0.73(66/90)	0.78(69/88)	0.76(135/178)
Model 2	17369	1486	786	700	0.71(67/95)	0.79(71/90)	0.75(138/185)
Model 3	15547	1629	737	892	0.69(62/91)	0.76(67/88)	0.72(129/179)
ABC3/0.90							
Model 1	18021	1016	642	374	0.82(62/76)	0.78(47/60)	0.80(109/136)
Model 2	17369	1129	617	512	**0.77(56/73)**	**0.86(62/72)**	**0.81(118/145)**
Model 3	15547	1311	624	687	0.83(63/76)	0.73(44/60)	0.79(107/136)
ABCH3/0.75							
Model 1	38797	3859	1790	2069	0.72(68/94)	0.76(68/90)	0.74(136/184)
Model 2	37636	4293	2007	2286	0.71(70/98)	0.76(74/97)	0.74(144/195)
Model 3	34407	4093	1785	2308	0.73(71/97)	0.65(62/95)	0.69(133/192)
ABCH3/0.90							
Model 1	38797	2746	1434	1312	0.76(63/83)	0.78(61/78)	0.77(124/161)
Model 2	37636	2923	1392	1531	0.75(63/84)	0.77(66/86)	0.76(129/170)
Model 3	34407	2949	1372	1577	0.74(66/89)	0.71(52/73)	0.73(118/162)

Footnotes: see Table 1

TABLE 3. Predictive performance of the CPDB mouse mammary carcinogen – nonmammary carcinogen SAR model with 3 to 7 heavy atoms.

Model	*Total. Fragments*	*Model Fragments*	*Active Fragments*	*Inactive Fragments*	*Sensitivity*	*Specificity*	*OCP#*
ABC3/0.75							
Model 1	5553	188	136	52	0.75(15/20)	0.61(11/18)	0.68(26/38)
Model 2	4718	138	69	69	0.80(16/20)	0.82(18/22)	0.81(34/42)
Model 3	6508	169	87	82	0.75(15/20)	0.78(14/18)	0.76(29/38)
ABC3/0.90							
Model 1	5553	106	73	33	0.80(12/15)	0.50(4/8)	0.70(16/23)
Model 2	4718	116	62	54	**0.79(15/19)**	**0.78(7/9)**	**0.79(22/28)**
Model 3	6508	122	69	53	0.83(15/18)	0.67(4/6)	0.79(19/24)
ABCH3/0.75							
Model 1	13517	801	591	210	0.62(13/21)	0.78(18/23)	0.71(31/44)
Model 2	12040	655	386	269	0.76(16/21)	0.82(18/22)	0.79(34/43)
Model 3	15187	753	434	319	0.62(13/21)	0.91(21/23)	0.77(34/44)
ABCH3/0.90							
Model 1	13517	443	324	119	0.55(11/20)	0.55(6/11)	0.55(17/31)
Model 2	12040	544	329	215	0.84(16/19)	0.74(14/19)	0.79(30/38)
Model 3	15187	553	352	201	0.79(15/19)	0.63(5/8)	0.74(20/27)

Footnotes: see table 1

TABLE 4. Predictive performance of the CPDB mouse mammary carcinogen – rodent noncarcinogen SAR model with 3 to 7 heavy atoms.

Model	*Total. Fragments*	*Model Fragments*	*Active Fragments*	*Inactive Fragments*	*Sensitivity*	*Specificity*	*OCP#*
ABC3/0.75							
Model 1	6414	379	72	307	0.84(16/19)	0.77(13/17)	0.81(29/36)
Model 2	6504	357	185	172	0.72(13/18)	0.65(11/17)	0.69(24/35)
Model 3	6157	294	172	122	0.75(12/16)	0.83(15/18)	0.79(27/34)
ABC3/0.90							
Model 1	6414	352	84	268	0.87(13/15)	0.44(4/9)	0.71(17/24)
Model 2	6504	244	192	52	**0.86(12/14)**	**0.40(4/10)**	**0.67(16/24)**
Model 3	6157	195	109	86	0.86(12/14)	0.75(6/8)	0.82(18/22)
ABCH3/0.75							
Model 1	14963	1396	436	960	0.85(17/20)	0.68(13/19)	0.77(30/39)
Model 2	15956	1502	672	830	0.63(12/19)	0.58(11/19)	0.61(23/38)
Model 3	14819	1188	658	530	0.79(15/19)	0.79(15/19)	0.79(30/38)
ABCH3/0.90							
Model 1	14963	1346	466	880	0.72(13/18)	0.80(12/15)	0.76(25/33)
Model 2	15956	1022	634	388	0.77(13/17)	0.65(11/17)	0.71(24/34)
Model 3	14819	1010	607	403	0.77(13/17)	0.67(10/15)	0.72(23/32)

Footnotes: see table 1

Appendices

Revised Statement of Work

The Statement of Work from the original application is provided below. The Statement of Work remains the same. We are requesting in the revised budget an extension on year one. To date, Specific Aim 1a and 1b are nearly complete and Specific Aim 1c has been started.

Title: Investigating the mechanisms of action and the identification of breast carcinogens by computational analysis of female rodent carcinogens

PI: Albert R. Cunningham, Ph.D.

Specific aim 1: Development and validation of SAR models for female breast carcinogens (months 1-12).

a. Identify chemicals tested in female rodents from the Carcinogenic Potency Database and the National Toxicology Program (month 1).

b. Enter chemical structures and potency values into MCASE program (months 2-8).

c. Validate models using 10-fold cross validation (months 9-12)

d. Summarize and interpret models and prepare publication.

Deliverable: If appropriate, publications describing rodent female and breast carcinogen models. This will include publishing data used to generate model and the achieved predictivity of the models for potential use in analyzing environmental chemicals for the identification of breast carcinogens.

Specific aim 2: Identify chemical and biological attributes of female and/or breast carcinogens to provide evidence to test the hypothesis that xenoestrogens are involved in breast cancer (months 13-36).

a. Compare and identify Structural Feature Overlap Method of female and breast carcinogens to those of other available toxicological SAR models (see Facilities and Equipment for a complete list of available models) (months 13-16).

b. As above using Joint Prevalence Method (months 16-24).

c. Identify the exact features of female and breast carcinogen models are responsible for predicted similar activities identified above (months 25-26).

d. Conduct QSAR and CoMFA analyses with chemicals containing these structures using biological data from appropriate assays (months 28-36).

e. Conduct metabolism experiments on identified outliers to see whether metabolic activation is required for activity and update models if required (months 28-36).

f. Summarize and interpret data and prepare publications (months 28-36).

Deliverables: Publication describing the role of estrogens and other toxicological events in the induction of female and breast cancer. Possible publication describing role of metabolism and biodegradation in converting inert chemicals into female or breast carcinogens. Publication assessing the overall accomplishments of using SAR analysis for the detection of breast cancer agents and the mechanistic information that was obtained describing the etiology of breast cancer and the mechanisms of action of breast carcinogens.

STATEMENT OF WORK UPDATE AUGUST 2004:

Due to a variety of reasons the PI has decided not to use MCASE for this project but to develop and use an alternative SAR program called cat-SAR. This has been discussed with Dr. Moore. The cat-SAR program has been developed and is capable of meeting the requirements of this project with the exception of Specific Aim 2e. This was not a significant Aim. We do, however, note that Professor

Rosenkranz at Florida Atlantic may have a working copy of MCASE in the near future thus allowing SA 2e to be pursued.

The reason for not using MCASE stems primarily from problems encountered in the cost of licensing. It approximately doubled since the original submission of the proposal. From a scientific point-of-view, the cat-SAR program stands alone from other computerized SAR expert systems in its openness, flexibility, routines for identifying important attributes of biological activity or inactivity, and its method for predicting the activity of untested compounds. Several commercially available computational SAR expert systems including MultiCASE, TOPKAT, and Oncologic are relatively closed systems where proprietory (and unknown) routines are used to generate the final model. On the other hand, cat-SAR is completely open with every detail of modeling transparent to the user. As for inflexibility, many of the commercially available expert systems maximally only allow the user to alter the makeup of the learning sets (users cannot alter the parameters for model development). The cat-SAR approach allows the user to select and/or adjust many parameters during the modeling process from learning set makeup, to selection of types of fragment attributes to consider, to ultimately what numerical or statistical considerations are employed in developing the final model.

Development of an information-intensive structure-activity relationship model and its application to human respiratory chemical sensitizers

A.R. Cunningham[1*], S.L. Cunningham[1], Daniel M. Consoer[1], Shanna T. Moss[1], and M.H. Karol[2]

[1]Department of Environmental Studies
Louisiana State University
Baton Rouge, LA 70803

[2]Department of Environmental and Occupational Health
University of Pittsburgh
Pittsburgh, PA 15261

Abbreviated Running Title: SAR of respiratory chemical sensitizers

*Corresponding author

ABSTRACT

Structure-activity relationship (SAR) models are recognized as powerful tools to predict the toxicologic potential of new or untested chemicals and also provide insight into possible mechanisms of toxicity. Models have been based on physicochemical attributes and structural features of chemicals. We describe herein the development of a new SAR modeling algorithm called cat-SAR that is capable of analyzing and predicting chemical activity from divergent biological response data. The cat-SAR program develops chemical fragment-based SAR models from categorical biological response data (e.g., toxicologically active and inactive compounds). The database selected for model development was a published set of chemicals documented to cause respiratory hypersensitivity in humans. Two models were generated that differed only in that one model included explicate hydrogen containing fragments. The predictive abilities of the models were tested using leave-one-out cross-validation tests. One model had a sensitivity of 0.94 and specificity of 0.87 yielding an overall correct prediction of 91%. The second model had a sensitivity of 0.89, specificity of 0.95 and overall correct prediction of 92%. The demonstrated predictive capabilities of the cat-SAR approach, together with its modeling flexibility and design transparency, suggest the potential for its widespread applicability to toxicity prediction and to deriving mechanistic insight into toxicologic effects.

Keywords:

structure-activity relationship (SAR); *in silico* modeling; respiratory sensitizer; predictive toxicology; chemical fragments; categorical SAR (cat-SAR) program

INTRODUCTION

The task of identifying toxic agents is not a small or trivial challenge. One approach has been to use mathematical models that relate biological activity to chemical structure. Benfenati and Gini [1] describe modern structure activity relationship (SAR) and quantitative SAR (QSAR) methods as typically involving three parts: 1) the chemical part, 2) the biological part (*i.e.*, activity), and 3) the methodology for relating parts 1 and 2. The main premise for these methods is that recurring and identifiable attributes of chemicals are associated with, or responsible for, particular biological effects. The attributes can take many forms including chemical structures, chemicophysical or quantum mechanical properties, and graph indices, to name a few. There are numerous methods that relate chemical structure with activity such as those based on human expertise like Ashby's "structural alerts" for potential carcinogenicity [2-4] to statistical QSAR methods like Hansch analysis (see [5]), comparative molecular field analyses (CoMFA) [6], and MCASE [7-9].

Advances in computing and chemoinformatics, standardized biological or toxicological testing, and the subsequent development of large libraries of test results have ushered in the era of computational or *in silico* SAR. Computational SAR models have gained recent acceptance in the regulatory community for both human health [10] and ecological endpoints [11]. Dearden succinctly summarized the field of computational SAR or *in silico* toxicity prediction to include QSAR models of congeneric and noncongeneric datasets and "expert systems" [12]. The utility and application of some important expert system toxicology prediction methods have been reviewed by Richard [13, 14]. Through the use of various techniques, the overall goal is to

identify meaningful associations between activity and chemical structure. These associations can then be used to investigate the underlying mechanisms of toxicity, or be extended to estimate or predict the toxicity of untested compounds.

With today's fast CPUs, abundant amounts of computer memory, and the availability of chemical informatics and graphics software we have aimed to readdress the challenge of computer-based SAR expert systems for modeling large and chemically diverse datasets. We describe herein the first generation of a new data and information-intensive approach to toxicological SAR modeling. The program is based on the well-established premise in SAR modeling that like structure begets like activity and employs chemical substructures to differentiate between categories of biologically active and inactive compounds for toxicological endpoints. We have named the new program cat-SAR for categorical SAR.

The cat-SAR program uses 2-dimensional chemical fragments generated by the Sybyl HQSAR module. We chose early in the development process of cat-SAR to use the Sybyl platform which already possessed the needed utilities of *in silico* chemical fragmenting, molecular graphics, and chemical informatics and database requirements associated with our modeling goals. Of importance, the HQSAR module is used solely to generate molecular fragments and is not used for further model development or statistical analysis.

Briefly, the HQSAR module is used to generate a list of chemical fragments associated with compounds in a learning set and produce a data matrix of compounds and fragments. In the data matrix, the rows are the chemicals and the columns are the molecular fragments. Thus for each

chemical, a tabulation of all its fragments are recorded across the table rows and for each fragment all chemicals that contain it are tabulated down the table columns. The compound-fragment matrix is then analyzed, in conjunction with the known biological activity category of each compound, by the cat-SAR program. The cat-SAR program identifies structural features associated with the biologically active and inactive categories. The cat-SAR program, the respiratory sensitizer learning set (described below), and the compound-fragment matrix are available through the corresponding author.

Since cat-SAR modeling is independent of the biological data used in the process we anticipate that it can be generally applied from the study of drugs to environmental toxicants. Moreover, the models can be used for either mechanistic studies of biological phenomena or for the prediction of biological activity for untested compounds.

The cat-SAR program stands alone from other computerized SAR expert systems in its openness, flexibility, routine for identifying important attributes of biological activity or inactivity, and its method for predicting the activity of untested compounds. Several commercially available computational SAR expert systems including MultiCASE, TOPKAT, and Oncologic are relatively closed systems where proprietory (and unknown) routines are used to generate the final model. On the other hand, cat-SAR is completely open with every detail of modeling transparent to the user. As for inflexibility, many of the commercially available expert systems maximally only allow the user to alter the makeup of the learning sets (users cannot alter the parameters for model development). The cat-SAR approach allows the user to select and/or adjust many parameters during the model process from learning set makeup, to selection of types

of fragment attributes to consider, to ultimately what numerical or statistical considerations are employed in developing the final model. These are described in detail below.

The cat-SAR approach is also a very data- and information-intensive SAR expert system. During model development and the creation of the final model, all fragments associated with the categories are presented. This leaves the user with an unbiased view of all important features associated with the biological endpoint. Consider the fact that the published MCASE model of the same respiratory sensitizer learning set used herein produced a model based on eight biophores and no biophobes [15]. One of the models developed with the cat-SAR program produced 1213 fragments associated with activity and 92 associated with inactivity. Similarly, the prediction of the activity of compounds outside the model's learning set presents the user with a *complete* correspondence between all the fragments in the model (e.g., 1213 active and 92 inactive) and those in the compound being predicted. Again considering the published MultiCASE report for this dataset, MultiCASE predicted the activity of methyldopa and presented the user with two reasons (i.e., biophores) for why the compound was predicted active. The cat-SAR program provided 22 reasons.

The approach we have taken in developing cat-SAR clearly diverges from existing SAR expert systems and is more in tune with modern QSAR techniques. For instance, the user is presented with a number of selectable and adjustable modeling parameters. The notion of having selectable and adjustable modeling parameters facilitates that ability to rigorously explore the relationships between chemical structure and biological activity.

We chose to test the method on a previously published respiratory sensitization model due to its small size (i.e., 80 compounds) and good modeling potential that was previously demonstrated using CASE-MultiCASE [15].

The cat-SAR program of course has some drawbacks and limitations. Like so many other expert systems in toxicology, it is applicable only to organic chemicals. Metals, mixtures, and polymeric compounds are not suitable for analysis. Moreover, as mentioned, the cat-SAR program presents the final SAR model, in terms of all relevant fragments. This lead to a model that may contain 1000s of fragments which may lead to difficulty in model interpretation.

This model has recently been reviewed by Rodford *et al.* [16]. Unlike other 2-dimensional modeling approaches including MultiCASE, the cat-SAR approach is transparent in development of the learning set, identification of fragments (i.e, biophores or activity descriptors), and determination of significant fragments. Moreover, the approach allows user intervention and model optimization throughout the modeling process. This method includes the ability to examine the entire fragment base, and to explore and optimize the fragments that have biological relevance.

MATERIALS AND METHODS

Description of the cat-SAR SAR Program

The cat-SAR models are built through a comparison of structural features found amongst the active and inactive compounds in the model's learning set. A categorical approach is used with, in this instance, compounds designated as active or inactive. For this exercise, active compounds were chemical respiratory sensitizers and inactive compounds were nonsensitizers. The modeling process began with the compilation of a set of chemicals and their biological activity (described below). Using the Tripos Sybyl HQSAR module, each chemical was fragmented into all possible fragments. HQSAR allows the user to select attributes for fragment determination including atom size, bond types, atomic connections, inclusion of hydrogen atoms, chirality and hydrogen bond donor and acceptor atoms. Moreover, fragments can be linear, branched or cyclic moieties.

We developed two sets of fragments from the model's learning set. The first (fragment set ABC) contained fragments between three and seven atoms in size and considered Atoms, Bonds types, and atomic Connections (i.e., the arrangement of atoms in the fragment). The second (fragment set ABCH) included the same descriptors as the previous set plus associated Hydrogen atoms. A compound-fragment matrix was produced for both sets of fragments.

A measure of each fragment's association with biological activity was next determined. This step is controlled by the user. To ascertain an association between each fragment and activity (or lack of activity) a set of rules is established to choose "important" active and inactive fragments. It should be noted that in this generation of the program we are using a common-sense approach, rather than statistical analysis, to select "significant" fragments.

The first selection rule is the number of times a fragment is identified in the learning set. For this exercise, it was arbitrarily set at three compounds (or 3.75%) of the compounds in the learning set. This was a reasonable decision considering that if a fragment is found in only one or two compounds in the learning set it may be a chance occurrence. We do, however, note that fragments found in only one or two compounds may not be outliers but rather underrepresented descriptors of activity. On the other hand, since the learning set is composed of only 40 active and 40 inactive compounds, if we required fragments to be found in more than three compounds, we would expect to miss important features.

The second rule relates to the proportion of active or inactive compounds that contain each fragment. For both the ABC and ABCH fragment sets, we set the proportion at 0.90. We reasoned that even if a particular fragment is associated with activity, there may yet be other reasons (i.e., fragments) for its being inactive, thus it would not be expected to be found in 100% of the active compounds. Likewise is true for inactive fragments. Thus, if we considered only those fragments found exclusively in active or inactive compounds we would rarify the fragments pool to an unreasonable level and risk losing valuable information. On the other hand, we expected that fragments found to be present approximately equally in the active and inactive fragment sets would not be associated with biological activity. Such fragments may serve as chemical scaffolds holding the biologically active features and are not directly related to activity or inactivity.

In summary, fragments were considered "significant" if they were found in at least three compounds in the learning set and also found in at least 90% of the active or inactive compounds that derived them. The two models developed are listed in Table I.

The resulting list of fragments can then be used for mechanistic analysis, or to predict the activity of an unknown compound. In the latter circumstance, the model determines which, if any, fragments from the model's learning set the compound contains. If none are present, no prediction of activity is made for the compound. If one or more fragments are present, the number of active and inactive compounds containing each fragment is determined. The probability of activity or inactivity is then calculated based on the total number of active and inactive compounds containing the fragments.

The probability of activity of a test chemical is calculated from the average probability of active and inactive fragments. For example, if a test compound contains two fragments, one is present 9/10 times in an active compound (i.e., 90% active) and one is found 3/3 times in an inactive compound (i.e., 100% inactive), the unknown compound will be predicted to be *inactive* based on the higher probability of inactivity derived from chemicals containing these fragments. In this manner, the probability of activity or inactivity is determined by comparison of the structure of the unknown compound with the entire structural information present in the model.

It requires noting that cat-SAR predictions are based on what can be conceived as two separable models: The inactive fragment model and the active fragment model. By so doing, cat-SAR predictions are based on information that is associated with biological activity and inactivity.

The cat-SAR program does not employ the use of default predictions wherein, as in the case of MultiCASE, if no biophores are present in an unknown chemical it is predicted by default to be inactive. This, of course, presents the situation wherein the cat-SAR program will not make predictions on some chemicals. Although this may seem like a drawback to the program by appearing less universal, the user of the program always has the option to simply define chemicals that are not predictable by cat-SAR with a default value.

Respiratory Sensitization Databases

The dataset of respiratory sensitizers has been reported [15]. Briefly, chemical sensitizers were identified through a search of the medical literature. Selection criteria were in accordance with the U.S. Department of Health and Human Services "Guidelines for Diagnosis and Treatment of Asthma" [17]. The search criteria included chemicals with inhalation challenge followed by a drop of >20% in forced expiration volume at 1 s within 24 h of challenge. Forty compounds were identified. No reports were identified of chemicals tested as described and found to be nonsensitizers in humans except for the often-used control substance, lactose. Since, as discussed, the cat-SAR method requires a comparison of biologically active with inactive compounds, we designated as "negative" a set of 40 chemicals previously selected as respiratory nonsensitizers by Graham *et al.* [15]. These 40 compounds were randomly selected from a dataset of chemicals tested for human allergic contact sensitizing ability via patch testing and were found to be nonsensitizers [18]. The assumption was made that dermal nonsensitizers

would also be respiratory nonsensitizers. In general, chemicals were relatively small organic compounds that did not include salts, metals, mixtures, or polymers.

RESULTS AND DISCUSSION

Predictive Performance of the cat-SAR Respiratory Sensitization Models

To evaluate the predictive ability of the models, a leave-one-out cross-validation test was conducted. For each chemical in the learning set, one at a time, its chemical fragments were removed from the total fragment set, and the probability of activity or inactivity associated with each fragment was recalculated. Using the criteria described above to estimate activity of unknown compounds, the activity of the removed chemical was predicted.

Overall, the ABC and ABCH models correctly classified 91% and 92% of the chemicals they were capable of predicting (Table I). The predicted activity for each chemical is listed in Table II. The cat-SAR program, using the n-1 leave-one-out cross-validation learning sets (i.e., models built on 79 compounds), was unable to make predictions for five chemicals in the ABC model and three in the ABCH (Table II). The reason for this is that each of these compounds did not possess any structural features that the n-1 models could base a prediction upon. A previous CASE/MultiCASE model of the same data reported an overall correct classification of 95%. This was based on the Bayesian combination of four CASE/MultiCASE submodels that individually had sensitivities ranging from 72% - 80% and specificities ranging from 95% - 98% [15]. In a separate published model based on chemicophysical parameters, a sensitivity of 85%

and a specificity of 74% was achieved [19]. Interestingly, the individual ABC and ABCH cat-SAR models are quite balanced with respect to sensitivity and specificity (Table I). This is not the case with the previous CASE/MultiCASE and chemicophysical models. The individual CASE/MultiCASE models tended to have a better ability to predict the inactive chemicals and the chemicophysical model was better able to predict the active ones.

The question arises as to why the program produced wrong predictions. In the case of any of the previously mentioned respiratory sensitizing models, the simplest explanation lies in the possibility that some of the information on which the models were built is not correct. Consider the National Toxicology Program's *Salmonella* mutagenicity database. The *Salmonella* database is derived from a standardized protocol and, more importantly, has been analyzed for reproducibility and accuracy by replicate analyses of chemicals [20]. The interlaboratory reproducibility of the *Salmonella* mutagenicity assay is only 85% [20]. Therefore, the databases may contain some incorrect information.

However, other explanations should be considered. The incorrect ABC model prediction for hexamethylene diisocyanate and the incorrect ABC and ABCH model predictions for isophorone diisocyanate are of interest. They both contain the isocyanate moiety which is clearly associated with biological activity. The cat-SAR program also identifies this moiety in these two compounds. However, the compounds contain a number of inactivating fragments that counterbalance the isocyanate-related ones. At this time, a complete understanding of the inaccurate predictions in not possible, but further development of both the models and the databases should lead to a more comprehensive analysis.

Respiratory Sensitization Model Analysis

As described above, two models were developed using the same set of 80 compounds. These models can be considered as independent since they are built upon different fragment bases. The ABC model started with a total fragment set of 5737 and the ABCH model with a set of 14424 fragments (Table I). In both models, approximately 23% of the total number of fragments met the criteria to be considered "significant" (i.e., 1307 significant /5753 total =22.7% for ABC and 3356 significant /144424 total =23.2%) (Table I). The remaining fragments were either not present in a sufficient number of compounds (i.e., found in <3 or 3.75% of compounds in the learning set), or the fragments did not come from compounds that were predominately (i.e., >90%) active or inactive.

Overall, both models performed similarly. However, when considering the sensitivity and specificity of the models, the distinction was not clear-cut. The ABC model was better able to correctly predict the active chemicals while the ABCH model was better able to predict the inactive ones. At this point, we chose to focus on the ABC model. This decision was based on several criteria: 1) Both models have nearly equivalent correct prediction rates (Table I) and make similar predictions on the majority of compounds in the validation set (Table II), 2) Considering the law of parsimony, the ABC model is based on fewer fragments, and 3) The models are constructed from a set of 40 chemicals *tested* and found to be respiratory sensitizers, whereas the set of 40 chemicals designated as "inactive" are *presumed* to lack activity. Therefore, based on the quality of information of these active and inactive sets, we favored a model with better ability to predict activity as compared with inactivity.

Although beyond the scope of this report, we bring attention to the finding that the cat-SAR method derives multiple independent models for the same endpoint. The observation that the ABC and ABCH models do not predict the same activity for each chemical suggests that the models may be capable of describing different attributes of the activity. This suggests the possibility of development of a consensus model using a Bayesian technique similar to those previously reported using CASE/MultiCASE [15].

Examples of the cat-SAR Model Predictions

Methyldopa and 2,4-dimethylbenzyl acetate were selected to demonstrate the predictive ability of the cat-SAR modeling method for an active and inactive chemical, respectively. For this demonstration, we used the ABC model for reasons just described. Tables III and IV list the significant fragments derived from the two compounds. Figures 1 and 2 illustrate the intact compounds and their associated fragments. The predictions presented for the two compounds are based on results obtained from the leave-one-out validation exercise. Therefore, the compounds themselves are not contributing to the fragment set of the model and are thus not influencing their own prediction of activity or inactivity.

Table III lists and Figure 1 shows all the significant fragments used in the leave-one-out validation exercise to predict the activity of methyldopa. Methyldopa was predicted to have a probability of activity of 0.988. This represents the average probability of activity of the 22

fragments used in the prediction (Table II). No fragments associated with methyldopa were considered inactive.

Likewise, Table IV and Figure 2 show all the significant fragments used in the validation exercise to predict the activity of 2,4-dimethylbenzyl acetate. 2,4-Dimethylbenzyl acetate was predicted to have a probability of inactivity of 1.0.

As indicated, the prediction for the respiratory sensitizing ability of methyldopa and 2,4-diemthylbenzyl acetate were based on the complete correspondence of significant fragments from the model's validation set to all the fragments identified in the compound. Methyldopa was predicted to be active based on 22 fragments from its validation set of fragments. Inspection of these fragments revealed several major themes. Fragment 348 leads to a series of complimentary moieties covering the amine to carboxylic acid portion of the molecule. Fragment 283 covers the *para* unsubstituted phenol and accounts for four other validation fragments. Fragment 2706 covers the 3,4-diol and accounts for five other validation fragments. Fragments 2415 and 2416 are closely related to Fragment 2706 but cover just the 3-hydroxyl.

For 2,4-dimethylbenzyl acetate, Fragments 4970 and 4979 cover the *para* substituted methyl section of the molecule. Moreover, Fragment 5073 covers the 2,4-methyl substitution and can account for four similar fragments.

From a prediction point-of-view, any one fragment would have been sufficient for the accurate prediction in these examples. From a mechanism point-of-view, for methyldopa, just the four

major fragment families (i.e., from fragments 348, 283, 2706, and 2416) would have covered the major identified structural themes relating to activity. The same is true for 2,4-dimethylbenzyl acetate where two sets of similar fragments (i.e., from fragments 5073 and 4970) described the compound. In this model, the fragment redundancy is obvious. However, we speculate that this may not be the case with other toxicological endpoints. In models for other endpoints, where fragments are similar but not exact, each fragment may contribute novel mechanistic and predictive information to the model.

Clearly, from the results of the validation exercises, the cat-SAR program in not performing at 100% accuracy. To judge the predictive performance of our models, we compared them to two previously developed MCASE models. One model is based on the National Toxicology Program's *Salmonella* mutagenicity database. The *Salmonella* database is derived from a standardized protocol and, more importantly, has been analyzed for reproducibility and accuracy by replicate analyses of chemicals [20]. The interlaboratory reproducibility of the *Salmonella* mutagenicity assay is 85% [20].

CONCLUSIONS

The new cat-SAR modeling approach described herein has a predictive ability in line with other respiratory sensitization models developed by us [15, 19]. This clearly suggests its utility and warrants further development. It is applicable to toxicological or pharmacological SAR modeling. The cat-SAR program uses a binary approach to identify structural features associated with biological activity or inactivity. This is straightforward when the toxicologic endpoint is

categorical (e.g., sensitizers *vs.* nonsensitizers, carcinogens *vs.* noncarcinogens or mutagens *vs.* nonmutagens). However, for other endpoints, where a continuous scale of activity is measured, the dichotomy can be imposed between highly active and less active compounds (e.g., extremely toxic *vs.* nontoxic as in the case of LD_{50} values or high or low receptor affinity as in the case of estrogen receptor ligands).

The cat-SAR method has two main areas of strength when compared with other 2-dimensional modeling systems. The first is the transparency of the method. The derivation of model fragments and decision rules are open for inspection. The entire compound-fragment matrix and the identified model fragments are all easily inspected. The second strength is the amount of user-selectable parameters available for adjustment. For the fragment development part of the program, the user can select fragments of different size and choose other fragment attributes including the consideration of atoms, bond, and hydrogen atoms. Moreover, when identifying important or significant fragments the user can manipulate the selection process by altering the requirements for how many compounds in the learning set contain each fragment and also what proportion of active or inactive compounds in the learning set contain the fragment.

Thus, the cat-SAR method is transparent with regard to the overall modeling process. Users of the program have the opportunity to optimize the process for their own needs. Considering the fact that toxicologic endpoints differ in their mechanisms, it makes sense that the modeling algorithm should be transparent to meet the requirements of the endpoint being modeled.

Overall, in prediction mode, this method presents the user with a *complete* correspondence of fragments in the model and the unknown chemical. In model analysis mode, the method provides the user with a complete listing of all interesting fragments. It should be noted that there is no hierarchy of fragments or filtering of "significant" fragments other than what the user chooses. There are no hidden or proprietary rules in the process. All fragments that meet the user-specified structural requirements and the rules of association with activity or inactivity are included in the model. This leads to the identification of many (e.g., 1000s) fragments, some with great structural similarity. This clearly presents difficulty in being able to succinctly describe the model. However, important information is retained and accessible to the user.

TABLE I Predictive performance of ABC and ABCH respiratory sensitization models. The ABC model was based on fragments of size between three and seven heavy atoms and considered atoms, bonds, and atom connection. The ABCH model also included consideration of hydrogen atoms.

Model	*Total. Fragments* *	*Model Fragments †*	*Active Fragments ‡*	*Inactive Fragments ¶*	*Sensitivity §*	*Specificity ‖*	*OCP #*
ABC	5737	1305	1213	92	0.94	0.87	0.91
ABCH	14424	3356	2926	430	0.89	0.95	0.92

Footnotes

* number of fragments derived from learning set.

† number of fragments meeting specified rules of the model.

‡ number of fragments meeting specified rules to be considered as active.

¶ number of fragments meeting specified rules to be considered as inactive.

§ number of correct positive predictions / total number of positives.

‖ number of correct negative predictions / total number of negatives.

Observed Correct Predictions: number of correct predictions / total number of predictions.

TABLE II Model validation for respiratory sensitizers. Compounds with values above 50% were predicted to be active compounds and those below 50% were predicted to be inactive.

		Model 3-7/3/0.90	
Chemical	*Experimental Activity*	*ABC % Active*	*ABCH % Active*
1,5-Napthalene diisocyanate	+	1.00	1.00
2-(*N*-Benzyl-*N*-*tert*-butylamino)-4'-hydroxy-3'-hydroxymethyl acetophenone diacetate	+	0.63	0.59
2,4-Toluene diisocyanate	+	1.00	1.00
2,6-Toluene diisocyanate	+	1.00	1.00
6-Amino penicillanic acid	+	1.00	1.00
7-Amino cephalosporanic acid	+	0.99	0.99
Ampicillin	+	1.00	1.00
Azocarbonamide	+	1.00	0.98
Benzylpenicillin	+	1.00	1.00
Brilliant orange GR	+	1.00	1.00
Carminic acid	+	0.57	0.54
Cephalexin	+	1.00	1.00
Chlorhexidine	+	1.00	0.96
Dichlorvos	+	*	*
Dimethyl ethanolamine	+	1.00	1.00
Diphenyl methane-4,4'-diisocyanate	+	1.00	1.00
Epigallocatechin gallate	+	0.57	0.60
Ethanolamine	+	1.00	1.00
Ethyl cyanoacrylate	+	*	0.03†
Ethylenediamine	+	1.00	1.00
Fenthion	+	0.91	0.96
Hexamethylene diisocyanate	+	1.00	0.38†
Isononanoyl oxybenzene sulfonate	+	0.98	0.82
Isophorone diisocyanate	+	0.22†	0.17†
Maleic anhydride	+	1.00	1.00
Methyl-2-cyanoacrylate	+	*	*
Methyldopa	+	0.99	0.95
Phenylglycine acid chloride	+	1.00	1.00
Phthalic anhydride	+	1.00	1.00
Piperacillin	+	1.00	1.00
Piperazine	+	1.00	1.00
Plicatic acid	+	0.53	0.74
Reactive orange 3R	+	1.00	1.00
Rifafix red BBN	+	1.00	1.00
Rifazol black GR	+	1.00	1.00
Tetrachloroisophthalonitrile	+	*	*
Tetrachlorophthalic anhydride	+	1.00	1.00
Triethylenetetramine	+	1.00	1.00
Trimellitic anhydride	+	1.00	1.00
Tylosin	+	0.14†	0.14†

1,1,3,3,5-Pentamethyl-4,6-Dinitroindane	-	0.00	0.00
1,4-Cineole	-	0.00	0.04
1-Hexanol	-	*	0.07
2,4-Dimethylbenzyl acetate	-	0.00	0.02
2-Butyl-4,4,6-trimethyl-1,3-dioxane	-	1.00†	0.50
2-*tert*-Amylcyclohexyl acetate	-	0.03	0.06
3,6-Dimethyloctan-3-yl acetate	-	0.05	0.06
3-Butyl phthalide	-	0.03	0.06
4-Acetyl-6-*tert*-butyl-1,1-dimethylindane	-	0.00	0.06
5-Methyl a-ionone	-	0.12	0.09
9-Decenyl acetate	-	0.05	0.05
Acetyl ethyltetramethyltetralin	-	0.00	0.00
Allyl heptylate	-	0.10	0.05
Benzyl butyrate	-	0.10	0.06
Butyl isobutyrate	-	0.06	0.07
Camphene	-	0.00	0.04
cis-3-Hexenyl anthranilate	-	0.65†	0.35
cis-4-Decen-1-al	-	0.03	0.04
Citronellyl nitrile	-	0.03	0.05
Cyclohexylethyl alcohol	-	0.00	0.06
Dibutyl sulphide	-	1.00†	0.93
Dihydro-isojasmone	-	0.03	0.04
Dimethylheptenol	-	0.03	0.05
Ethyl acetoacetate ethylene glycol ketal	-	0.27	0.19
Ethyl lactate	-	0.09	0.07
Eugenyl phenylacetate	-	1.00†	0.81†
?-Dodecalactone	-	0.05	0.07
Geranyl benzoate	-	0.03	0.06
Heptyl butyrate	-	0.06	0.06
Hexane	-	0.00	0.09
Hexyl tiglate	-	0.04	0.06
Isoamyl butyrate	-	0.06	0.06
Lactoscatone	-	0.04	0.05
l-Carvyl propionate	-	0.04	0.04
Methyl tiglate	-	0.09	0.07
Musk xylol	-	0.00	0.00
Phenylethyl acetate	-	0.77†	0.32
p-Isopropylcyclohexanol	-	0.00	0.04
Rhodinyl formate	-	0.03	0.05
Undecenyl acetate	-	0.05	0.05

Footnotes

* no prediction was made for the compound

† wrong prediction was made for the compound

TABLE III Fragments from the ABC model leave-one-out validation analysis used to predict the activity of the respiratory sensitizer methyldopa.

Fragment	*No. Active**	*No. Inactive†*	*Total‡*	*% Active*	*% Inactive*
frag258	10	1	11	0.909	0.091
frag283	10	1	11	0.909	0.091
frag308	10	1	11	0.909	0.091
frag348	8	0	8	1.000	0.000
frag357	8	0	8	1.000	0.000
frag400	14	0	14	1.000	0.000
frag471	6	0	6	1.000	0.000
frag522	6	0	6	1.000	0.000
frag914	4	0	4	1.000	0.000
frag915	4	0	4	1.000	0.000
frag920	4	0	4	1.000	0.000
frag921	4	0	4	1.000	0.000
frag2378	3	0	3	1.000	0.000
frag2401	3	0	3	1.000	0.000
frag2415	3	0	3	1.000	0.000
frag2416	3	0	3	1.000	0.000
frag2463	3	0	3	1.000	0.000
frag2471	3	0	3	1.000	0.000
frag2472	3	0	3	1.000	0.000
frag2507	3	0	3	1.000	0.000
frag2509	3	0	3	1.000	0.000
frag2706	3	0	3	1.000	0.000
Probability of activity				0.988	0.012

Footnotes:

* number of active compounds that contain the fragment

† number if inactive compounds that contain the fragment

‡ number of compounds in the dataset that contain the fragment

TABLE IV Fragments from the ABC model leave-one-out validation analysis used to predict the activity of the respiratory nonsensitizers 2,4-Dimethylbenzyl acetate.

Fragment	*No. Active**	*No. Inactive†*	*Total‡*	*% Active*	*% Inactive*
frag4970	0	3	3	0.000	1.000
frag4979	0	3	3	0.000	1.000
frag4982	0	3	3	0.000	1.000
frag5003	0	4	4	0.000	1.000
frag5011	0	4	4	0.000	1.000
frag5032	0	4	4	0.000	1.000
frag5033	0	4	4	0.000	1.000
frag5073	0	4	4	0.000	1.000
	Probability of activity			0.000	1.000

Footnotes: See Table III

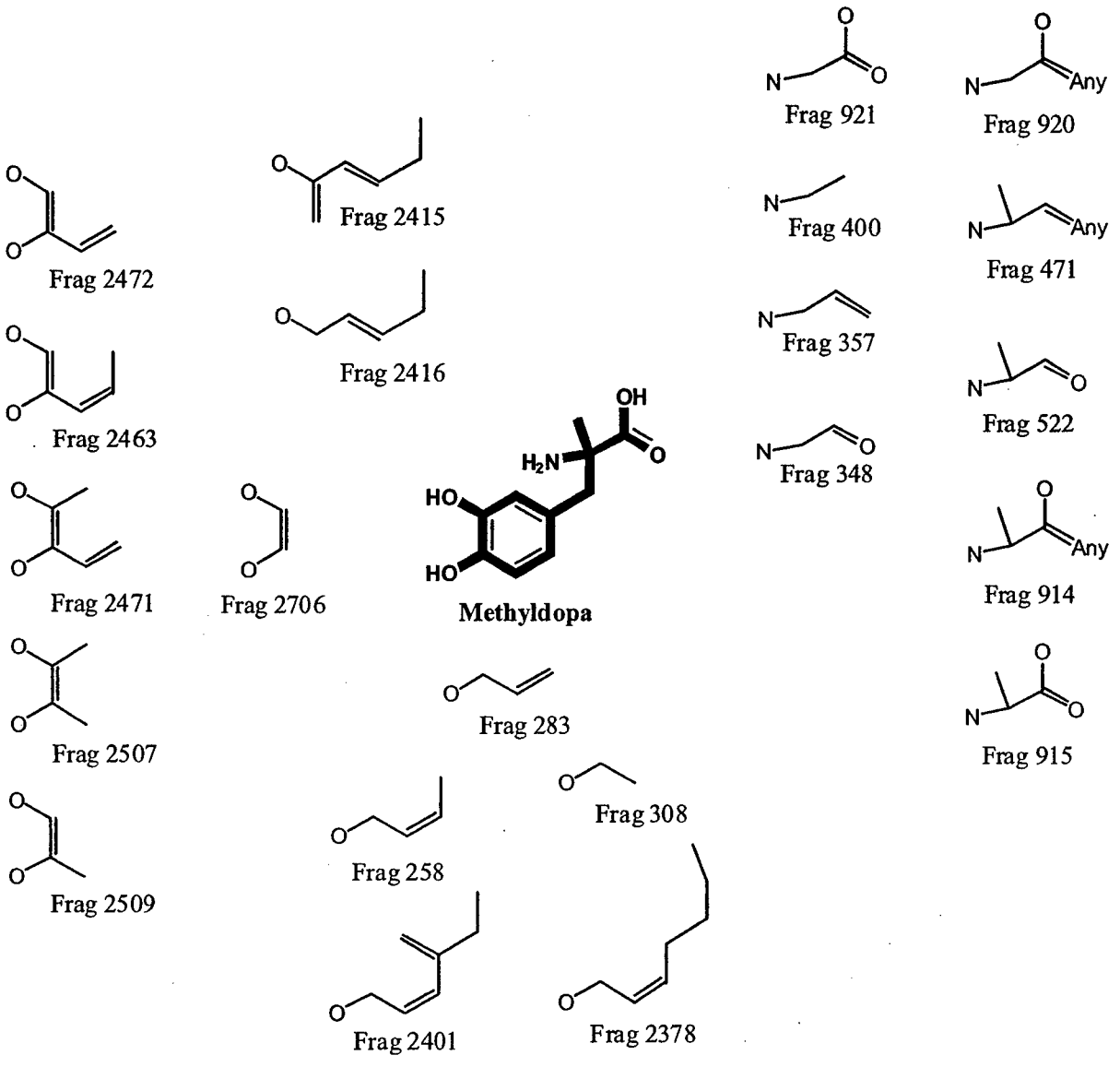
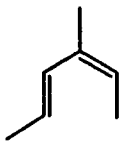
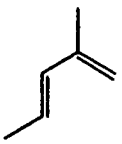


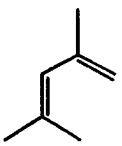
FIGURE 1 Illustration of the 22 significant fragments contributing to the active validation prediction of methyldopa.



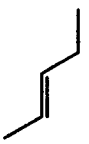
Frag 5011



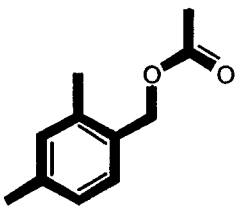
Frag 5033



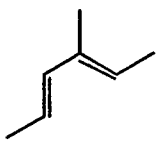
Frag 5032



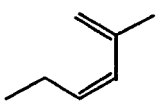
Frag 5073



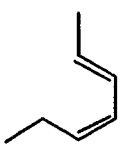
2,4-diemthylbenzyl acetate



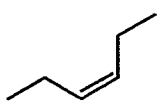
Frag 4982



Frag 4970



Frag 5003



Frag 4979

FIGURE 2 Illustration of the eight significant fragments contributing to the inactive validation prediction of 2,4-dimethylbenzyl acetate.

acknowledgments

We gratefully acknowledge support for the development of the cat-SAR program from the Department of Defense Congressionally Directed Medical Research Program for Breast Cancer Idea Award DAMD17-01-0376.

references

[1] Benfenati, E. and Gini, G. (1997) "Computational predictive programs (expert systems) in toxicology", *Toxicology* **119,** 213-225.

[2] Ashby, J. and Paton, D. (1993) "The influence of chemical structure on the extent and sites of carcinogenesis for 522 rodent carcinogens and 55 different human carcinogen exposures", *Mutat. Res.* **286,** 3-74.

[3] Ashby, J. (1985) "Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity", *Environ. Mutagen.* **7,** 919-921.

[4] Ashby, J. and Tennant, R.W. (1991) "Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP", *Mutat. Res.* **257,** 229-306.

[5] Hansch, C. and Leo, A. 1995 Exploring QSAR Fundamentals and Applications in Chemistry and Biology. (American Chemical Society, Washington, D.C.).

[6] Cramer, R.D., Patterson, D.E., and Bunce, J.D. (1988) "Comparative molecular field analysis (CoMFA). 1. Effects of shape on binding of steroids to carrier proteins", *J. Am. Chem. Soc.* **110,** 5959-5967.

[7] Klopman, G. (1984) "Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules", *J. Am. Chem. Soc.* **106,** 7315-7321.

[8] Klopman, G. (1992) "MULTICASE 1. A hierarchical computer automated structure evaluation program", *Quant. Struct. Act. Relat.* **11,** 176-184.

[9] Klopman, G. and Rosenkranz, H.S. (1994) "Approaches to SAR in carcinogenesis and mutagenesis. Prediction of carcinogenicity / mutagenicity using MULTI-CASE", *Mutat. Res.* **305,** 33-46.

[10] Cronin, M.T.D., Jaworska, J.S., Walker, J.D., Comber, M.H.I., Watts, C.D., and Worth, A.P. (2003) "Use of quantitative structure-activity relationships in international decision-making frameworks to predict health effects of chemical substances", *Environ. Health Perspect.* **111,** 1376-1390.

[11] Cronin, M.T.D., Walker, J.D., Jaworska, J.S., Comber, M.H.I., Watts, C.D., and Worth, A.P. (2003) "Use of quantitative structure-activity relationships in international decision-making frameworks to predict ecological effects and environmental fate of chemical substances", *Environ. Health Perspect.* **111,** 1376-1390.

[12] Dearden, J.C. (2003) "*In silico* prediction of drug toxicity", *J. Comput. Aided Mol. Des.* **17,** 119-127.

[13] Richard, A.M. (1998) "Commercial toxicology prediction systems: a regulatory perspective", *Toxicol. Lett.* **102-103,** 611-616.

[14] Richard, A.M. (1999) "Application of artificial intelligence and computer-based methods to predicting chemical toxicity", *Knowl. Eng. Rev.* **14,** 307-317.

[15] Graham, C., Rosenkranz, H.S., and Karol, M.H. (1997) "Structure-activity model of chemicals that cause human respiratory sensitization", *Regul. Toxicol. Pharmacol.* **26,** 296-306.

[16] Rodford, R., Patlewicz, G., Walker, J.D., and Payne, M.P. (2003) "Quantitative structure-activity relationships for predicting skin and respiratory sensitization", *Environ. Toxicol. Chem.* **22,** 1855-61.

[17] USDHHS. (1991) "Guidelines for Diagnosis and Treatment of Asthma", *U.S. Department of Health and Human Services, National Institutes of Health, Publication No. 90-3042.*

[18] Graham, C., Gealy, R., Macina, O.T., Karol, M.H., and Rosenkranz, H.S. (1996) "QSAR for allergic contact dermatitis", *Quant. Struct. Act. Relat.* **15,** 224-229.

[19] Karol, M.H., Macina, O.T., and Cunningham, A.R. (2001) "Cell and molecular biology of chemical allergy", *Ann. Allergy. Asthma. Immunol.* **87,** 28-32.

[20] Piegorsrch, W.W. and Zeiger, E. (1991) "Measuring Intra-assay Agreement for the Ames *Salmonella* Assay". In: Hotham, L., ed) Statistical Methods in Toxicology. (Springer-Verlag, Heidlberg), pp. 35-41.

Figure Titles Page

FIGURE 1 Illustration of the 22 significant fragments contributing to the active validation prediction of methyldopa.

FIGURE 2 Illustration of the eight significant fragments contributing to the inactive validation prediction of 2,4-dimethylbenzyl acetate.

SAR and QSAR in Environmental Research, VOL. 15 (1), February 2004, pp. 55–67

Taylor & Francis
Taylor & Francis Group

STRUCTURE–ACTIVITY APPROACH TO THE IDENTIFICATION OF ENVIRONMENTAL ESTROGENS: THE MCASE APPROACH

A.R. CUNNINGHAM[a,*], S.L. CUNNINGHAM[a] and H.S. ROSENKRANZ[b]

[a]*Department of Environmental Studies, Louisiana State University, 1285 Energy, Coast & Environment Building, Baton Rouge, LA 70803 USA;* [b]*Department of Biomedical Sciences, Florida Atlantic University, Boca Raton, FL 33431 USA*

(Received 24 April 2003; In final form 25 October 2003)

A sizable number of environmental contaminants and natural products have been found to possess hormonal activity and have been termed endocrine-disrupting chemicals. Due to the vast number (estimated at about 58,000) of environmental contaminants, their potential to adversely affect the endocrine system, and the paucity of health effects data associated with them, the U.S. Congress was led to mandate testing of these compounds for endocrine-disrupting ability. Here we provide evidence that a computational structure–activity relationship (SAR) approach has the potential to rapidly and cost effectively screen and prioritize these compounds for further testing. Our models were based on data for 122 compounds assayed for estrogenicity in the ESCREEN assay. We produced two models, one for relative proliferative effect (RPE) and one for relative proliferative potency (RPP) for chemicals as compared to the effects and potency of 17β-estradiol. The RPE and RPP models achieved an 88 and 72% accurate prediction rate, respectively, for compounds not in the learning sets. The good predictive ability of these models and their basis on simple to understand 2-D molecular fragments indicates their potential usefulness in computational screening methods for environmental estrogens.

Keywords: Environmental estrogens; Xenoestrogens; Structure–activity relationship (SAR); Computational modeling

INTRODUCTION

Compounds that mimic the activity of 17β-estradiol are of interest and concern for several reasons. First, many environmental contaminants have been found to possess estrogenic activity. These xenoestrogens are more generally known as endocrine disruptors. Another group of estrogenically active agents are of medicinal value. These are the selective estrogen receptor modulators (i.e. SERMs) that are actively being investigated as breast cancer therapies. The widely used tamoxifen and to a lesser extent rolaxifene are two such examples. Additionally, interest is focusing on plant derived estrogens (i.e. phytoestrogens) as chemopreventative agents [1] as well as alternative therapies for postmenopausal hormone replacement therapies [2,3].

*Corresponding author. E-mail: arc@lsu.edu

ISSN 1062-936X print/ISSN 1029-046X online © 2004 Taylor & Francis Ltd
DOI: 10.1080/1062936032000169679

With the obvious usefulness of SERMs, medicinal chemistry has added a great deal of understanding to the phenomena of estrogenicity and some of the health effects associated with these compounds. Although exceptionally useful, the investigation of SERMs does not cover the entire plethora of environmental concerns regarding endocrine active agents. The consequence of exposure to estrogen mimics can cause a vast array of toxicological and pharmacological responses including cancer [4–6], cancer therapy, developmental abnormalities and altered sexual differentiation [7,8], immune disturbances [9] as well as no observable adverse effects or even beneficial responses [1]. It has also been observed that the timing (e.g. fetal vs. adult), hormonal status, and level and duration of exposure affect the biological consequences associated with exposure to these agents. Moreover, apart from diversity in biological response, the estrogen mimics, as a group, display minimal structural homology [10]. This presents a challenge to structure–activity relationship (SAR) approaches aimed at their identification (i.e. predicted activity), activity and understanding mechanisms of action.

The United States Environmental Protection Agency (EPA) was mandated under the 1996 Food Quality Protection Act by the United States Congress to develop a screening and testing strategy to determine whether exogenous substances may have an effect in humans similar to those of natural hormones [11]. The EPA considers 87,000 chemicals as potentially requiring analysis for endocrine activity [12]. To facilitate this, a stated key goal of the EPA is to pursue computational methods for their analysis [13]. Computational SAR have gained recent acceptance in the regulatory community for both human health [14] and ecological endpoints [15].

Waller and others [16–19] have demonstrated the ability of comparative molecular field analysis (CoMFA) to accurately predict the relative binding affinities (RBA) of several series of compounds for the estrogen receptor. However, due to the limitations on CoMFA, these analyses had to rely on congeneric series of compounds for the training sets. However, with this limitation, these models are quite capable of predicting the activity of compounds that fit this model space. Additionally, the National Center for Toxicological Research has published a set of rat uterine cytosol RBA data [20]. Shi *et al.* [21] successfully analyzed this dataset and produced predictive CoMFA and holographic quantitative structure–activity relationship (HQSAR) models. Moreover, this same group has recently demonstrated the use of structural alerts for estrogen activity in a logical tree-based method to prioritize upwards of 58,000 compounds that are of environmental concern [22].

The ESCREEN dataset was chosen for several reasons. Basically, the ESCREEN assay measures estrogen-induced growth of human MCF-7 breast cancer cells [23,24]. Given the broad spectrum of biological assays for estrogenicity, the ESCREEN assays fall somewhere in the middle of the biological complexity scale (i.e. above *in vitro* receptor binding and below *in vivo* whole animal assays). This assay is well characterized, and the investigators report estrogenic response of chemicals using two unique parameters (i.e. relative proliferative potency (RPP) and relative proliferative effect (RPE)). RPP is the ratio between the least amount of 17β-estradiol needed to produce maximum proliferation and the least amount of the test chemical needed to produce a comparable effect [25]. That is, RPP compares the estrogenic potency of a compound to the potency of the standard estrogen 17β-estradiol. On the other hand, it is realized that many estrogenic compounds, no matter how high the dose, will never produce cell proliferation at the rate of 17β-estradiol. The RPE measures this effect. The PRE is 100 times the ratio of the greatest cell yield obtained with a test chemical and that obtained by 17β-estradiol [25].

The present investigation uses the MCASE algorithm (MultiCASE, Inc., Beechwood, OH) to predict estrogenic activity as measured in the ESCREEN assay [25]. The advantage of this approach is its ability to deal with non-congeneric datasets, as does HQSAR.

However, unlike HQSAR and CoMFA approaches that require continuous-type data, MCASE works by identifying molecular attributes associated with biological activity by comparing attributes of active (i.e. estrogenic) to inactive (i.e. non-estrogenic) compounds (i.e. a binary-type response). Although the MCASE program uses binary information to discriminate among structural features associated with active and inactive compounds, the program in this setting also took into account potency values for the active compounds. The models and subsequent predictions based on this dichotomy can then be used to examine structural features associated with estrogenicity and predicted the potential estrogenic activity of unknown compounds respectively.

The present report demonstrates the ability of MCASE to adequately assess compounds for their ability to induce an estrogenic response in MCF-7 cells. With these promising results, we are currently assessing the method's applicability to assess estrogen receptor binding ability as well as uterine growth stimulation and inhibition. Overall, considering the work from the National Center for Toxicological Research and the preliminary SAR modeling approach discussed here for environmental estrogens, it seems plausible that computational methods singly or in combination will be able to provide a reliable method to prioritize compounds for further testing and for regulatory classification. These methods, could therefore drastically reduce the tremendous financial cost, time and use of animals associated with meeting the mandate to assess these compounds for endocrine disrupting ability.

MATERIALS AND METHODS

Database

Two learning sets of 122 chemicals were created from publications of Soto and colleagues [23,25]. Both sets contain the same chemicals. The RPP learning set consisted of 50 active (i.e. estrogenic) and 72 inactive (i.e. non-estrogenic) chemicals. The RPE learning set consisted of 73 active and 49 inactive chemicals. Potency values for each endpoint were scaled to conform to MCASE requirements that SAR potency units range between 10 and 99 activity units. In this scale, inactive compounds were less than 30 and actives were greater than or equal to 30. Inactive RPE compounds were assigned 10 units and active compounds were scaled using the conversion equation, SAR units $= 0.62\,(\text{RPE}) + 29.38$. Inactive RPP compounds were assigned 10 SAR activity units and the active compounds were scaled using the conversion equation, SAR units $= 9.57 \log\,(\text{RPP}) + 70.29$.

The potency values obviously differed between the RPP and RPE models. However, the overall designation of compounds as estrogenic or non-estrogenic also differed between the two. Twenty three chemicals designated as inactive in the RPP set were listed as active in the RPE model (compounds 2,2′,3,3′,5,5′-hexachlorobiphenyl through 6-bromonaphthol-2, Table II). All the 23 compounds in question had very low RPE values. Although the original authors of the studies chose to call these compounds non-estrogens, we chose to call them active since activity (although minimal) was reported.

MCASE Methodology

The MCASE methodologies have been described [26–28]. Basically, MCASE selects its own descriptors automatically from a learning set composed of active and inactive molecules. The descriptors are readily recognizable single, continuous structural fragments that are embedded in the complete molecule. The descriptors consist of either activating or inactivating fragments termed as biophores and biophobes, respectively. Each of

the fragments is associated with a confidence level and a probability of activity that is derived from the distribution of these biophores and biophobes among active and inactive molecules. MCASE then selects the most important of these fragments as a biophore (i.e. the functionality that is associated with the largest number of active molecules and fewest number of inactive molecules). A biophore may also be a 2-D distance descriptor based upon the presence of lipophilic centers or heteroatoms in the molecule [29]. At this point, a congeneric series of chemicals has been identified with the biophore being the unifying feature. MCASE then performs a series of defined chemical substitutions of the atoms in the first biophore (e.g. halogen for halogen or nitrogen for carbon in aromatic systems) and then searches for similar biophores in the pool of fragments significantly related to active chemicals. All chemicals containing these related structural features are grouped together under a single biophore designation. Thus, a biophore may consist of a single feature or a family of chemically similar features. Using the molecules contained in this family as a learning set, MCASE derives a local QSAR equation for this series of chemicals. The regression variables may be chemical properties (e.g. structural fragments), physicochemical (e.g. $\log P$, water solubility), or quantum chemical parameters such as the energy of the highest occupied molecular orbital (HOMO) and the energy of the lowest unoccupied molecular orbital (LUMO). These features ("modulators") thus augment or decrease the basal activity associated with the biophore. The identified biophore and modulators will then be used to derive a local QSAR equation for chemicals within this subset. If the data set is congeneric, then the single biophore and associated modulators may explain the activity of the entire training set; this usually does not occur and there is a residue of molecules not explained by the single biophore and modulators. When this happens, the program will remove from consideration the molecules already explained by this biophore and will search for the next biophore and associated modulators. The process is iterated until all of the active molecules in the learning set have been explained or until no further biophores are identified.

The MCASE SAR program yields two numerical parameters when challenged with unknown chemicals. These are a predicted probability of activity and a predicted potency value. We have found that the ability to identify active or inactive compounds can be optimized by separate analyses of each of the two parameters to define optimal cutoff values for each that best separate predicted active from predicted inactive chemicals and therefore yield the best concordance between predictions and experimental results. Bayes' Theorem was used to combine the two individual parameters to yield an indication of the model's overall sensitivity, specificity and concordance [30,31] (Table I). Briefly, Bayes' Theorem

TABLE I Predictive performance summary for RPP and RPE MCASE models

Model	*Concordance*	*Sensitivity*	*Specificity*
RPP			
SAR Units	0.72	0.72	0.72
Probability	0.74	0.70	0.76
Overall	0.72	0.72	0.72
RPE			
SAR Units	0.87	0.88	0.86
Probability	0.86	0.92	0.78
Overall	0.88	0.86	0.89

Notes:
Concordance: number of correct predictions / total number of predictions.
Sensitivity: number of correct positive predictions / total number of positives.
Specificity: number of correct negative predictions / total number of negatives.
Overall: combined SAR models derived from Bayes' Theorem.

states that the joint probability of two events is the product of the probability of one of the events and the conditional probability of the second event, given that the first event occurs. The system employed here starts with a prior probability for the first event, which is set at 0.5. This reflects the fact that SAR models are constructed wherein the ratio of active to inactive chemicals is unity. Using Bayes' Theorem, this prior probability (i.e. 0.5) is updated with the specificity and sensitivity of the first SAR submodel. This posterior probability serves as the new prior probability to which the sensitivity and specificity of the second submodel is incorporated. This process is iterated for the two MCASE parameters to derive an overall probability that a chemical is active based upon the combined information of the SAR submodels [30–32].

To examine the predictivity of the MCASE SAR models, 10-fold cross-validation tests were conducted [33]. From the RPE and RPP learning sets, 10 mutually exclusive test sets were prepared. These sets were created by the random removal of approximately 10% of the chemicals in the database. The activity of each chemical in the test set was predicted from models developed with the remaining 90% of the database as a learning set. This allowed the determination of sensitivity, specificity and concordance between experimental and predicted results.

To analyze the potential of chemicals demonstrating estrogenic activity in the ESCREEN assay to induce other toxicological phenomena including cancer and developmental toxicity we used the "Chemical Diversity Approach". This is a method based on comparisons of the predicted toxicological profiles of a group of 10,000 chemicals chosen to represent a random assortment of all chemicals and chemical features [34]. These chemicals were derived from chemical structure libraries and from a random sample of chemical structures from the National Cancer Institute Repository of potential cancer chemotherapeutic agents. The various toxicological properties of these chemicals are predicted using validated SAR models including the models for RPE and RPP. The prevalence of chemicals predicted to possess two toxicological properties simultaneously is then quantified and compared to the expected prevalence. If the two effects are assumed to be independent of one another (i.e. null hypothesis), then the observed and expected values should be nearly equal. A significantly greater observed than expected prevalence indicates a similarity in mechanism among the toxicological effects that are being studied. Likewise, a significantly lower observed than expected prevalence suggests a possible antagonism between the phenomena under investigation. The applicability of the methodology to the study of diverse toxicological phenomena has been demonstrated by successfully estimating the number of potential *Salmonella* mutagens in the environment [35]. The inhibition of gap junctional intercellular communication is related to rodent carcinogenesis through cellular and systemic toxicity but not genotoxicity [34].

RESULTS AND DISCUSSION

Examination of the performance of the RPE and RPP models indicates that both have acceptable predictive performances to identify estrogenic and non-estrogenic compounds (Tables I and II). Overall, the RPP model correctly assessed the estrogenic activity of 72% of the compounds not included in the learning sets, while the RPE model correctly assessed 88% of the compounds. Interestingly, using the same 122 compounds, the RPE model outperformed the RPP models by 16%. This was achieved by both an increase in sensitivity and specificity. The change of 23 activity designations (see above) could have altered the structural components of the models to the point that they were over-weighted with either active or inactive chemicals and thus possibly, over-predict either group. However, this

was not the case as each overall model nearly equally maintained values for sensitivity and specificity (Table I). Therefore, the results suggest that these exceedingly weak chemicals are nonetheless true estrogens and contribute information to the model.

As mentioned, 23 compounds (compounds 2,2′,3,3′,5,5′-hexachlorobiphenyl through 6-bromonaphthol-2, Table II) have disparate activity in the two assays. These were mostly weak RPE compounds and negative RPP ones which were in some instances categorized as negative. Interestingly, 22 of the 23 are accurately predicted for RPE activity (Table II). However, the RPP model "erroneously" predicted these RPP inactive compounds as active. The predictivity of a model has been used as an acceptable measure for assessing the "meaningfulness" of a model [36]. Moreover, we have consistently observed that good predictivity is based on mechanistically sound and interpretable models [37,38]. Therefore, we consider that the RPE model, which includes the very weak estrogens, is a more informative model than that based on the RPP data set. This finding is significant with respect to applying the model to environmental estrogens and phytoestrogens, many of which are exceedingly weak compared to 17β-estradiol.

The major structural attributes that composed both models are listed in Tables III and IV and shown in Figs. 1 and 2. The sets of biophores designated with the letter are expanded biophores. Once the primary biophore (i.e. version a) is identified, the program searches for similar structural fragments to include in the expanded biophore family. The MCASE model for RPE consisted of five biophores (Table III and Fig. 1) and the model for RPP consisted of nine biophores (Table IV and Fig. 2). The RPE biophores divided the data into three basic groups: a phenolic ring with varying substitution patterns, the chlorinated nonaromatic compounds, and moieties that depict the keto and hydroxy substituents of 17β-estradiol derivatives. The RPP model was made up of biophores that were similar in nature to the RPE biophores. The major biophore in each model was the phenolic A-ring. It should be noted that this major biophore although not specific for a hydroxyl substitution which is commonly associated with estrogenicity was derived predominately from phenols.

Interestingly, the RPE biophores were more robust, each typically being derived from more chemicals than those in the RPP model. For example, what is explained with biophores 1–4 in the RPP model is explained with only two biophores in the RPE model. It is noteworthy that the RPE model outperformed the RPP model with fewer structural moieties being associated with activity. Therefore, designating the 23 weak estrogens as active compounds facilitated a refinement of features associated with estrogenicity. That is to say, the model contained more robust structural features and thus also indicates the superiority of the RPE over the RPP model.

Since the RPE model is both simpler and more predictive, we used it in the "Chemical Diversity Approach" to investigate the possible role of estrogens in other toxicological phenomena. Essentially, based on the greater than expected prevalence of chemicals possessing two toxicological properties simultaneously (one being estrogenicity in this exercise) we can hypothesize on the underlying mechanisms of action being related. The first analysis consisted of comparing the RPP and RPE models. As expected, there was a high degree of similarity verifying that, although they both are measuring different estrogenic endpoints (i.e. proliferative potency and effect relative to 17β-estradiol), these endpoints are related (Table V, Analysis 1). The two major health concerns related to environmental estrogens are their potential to induce cancer and developmental effects. We found that generally the RPE model did not significantly overlap with chemicals that have the potential to induce mutagenicity, unscheduled DNA synthesis, and chromosomal aberrations (Table V, Analyses 2–4). This is not unexpected as estrogens are not genotoxic *per se*. Only the SOS Chromotest showed significant commonality with estrogens (Table V, Analysis 5). This may be a reflection of the fact that this assay, unlike the others, responds

TABLE II Experimental results and MCASE predictions for RPP and RPE

	RPP		*RPE*	
Chemical	*Experimental*	*Prediction*	*Experimental*	*Prediction*
1,2-Dichloropropane	-	-	-	-
1-Naphthol	-	+	-	-
2,3,7,8-TCDD	-	-	-	-
2,4-DB Acid	-	-	-	-
2,4-Dichlorophenoxyacetic acid	-	-	-	-
2-Naphthol	-	+	-	+
4-Butooxyphenol	-	+	-	-
4-Hexyloxyphenol	-	+	-	-
5,6,7,8-Tetrahydronaphthol-2	-	+	-	-
Alachlor	-	-	-	-
Atrazine	-	-	-	-
Bendiocarb	-	-	-	-
Butylate	-	-	-	-
Butylated hydroxytoluene	-	-	-	-
Carbaryl	-	-	-	-
Carbofuran	-	-	-	-
Chlordimeform	-	-	-	+
Chlorothalonil	-	-	-	-
Chlorpyrifos	-	-	-	-
Cyanazine	-	-	-	-
Dacthal	-	-	-	-
Diamyl phthalate	-	-	-	-
Diazinon	-	-	-	-
Dibutyl phthalate	-	-	-	-
Dimethyl isophthalate	-	-	-	-
Dimethyl terephthalate	-	-	-	-
Dinonyl phthalate	-	-	-	-
Dinoseb	-	-	-	-
Hexachlorobenzene	-	-	-	-
Hexazinone	-	-	-	-
Kelthane	-	-	-	+
Lindane	-	-	-	-
Malathion	-	-	-	-
Maneb or zineb	-	-	-	-
Methoprene	-	-	-	-
Metalochlor	-	-	-	-
Mirex	-	+	-	+
Octachlorostyrene	-	-	-	-
Parathion	-	+	-	-
Phenol	-	+	-	-
Picloram	-	-	-	-
Propazin	-	-	-	-
Rotenone	-	-	-	-
Simazine	-	-	-	-
Styrene	-	-	-	+
Tetrachloroethylene	-	-	-	-
Thiram	-	-	-	-
Trifluralin	-	-	-	-
Ziram	-	-	-	-
2,2′,3,3′,5,5′-Hexachlorobiphenyl	-	-	1	+
2,3,3′,4,5-Pentachlorobiphenyl	-	+	1	+
3,5-Dichloro-4-hydroxybiphenyl	-	-	1.5	+
4-Monochlorobiphenyl	-	-	2.1	+
2,3′,5-Trichlorobiphenyl	-	-	2.2	+
3,5-Dichlorobiphenyl	-	-	2.7	+
2,3,5,6-Tetrachlorobiphenyl	-	-	3.1	+
2,6-Dichlorobiphenyl	-	-	3.4	+
Decachlorobiphenyl	-	-	3.5	+
2,5-Dichlorobiphenyl	-	-	3.7	+
Chlordene	-	+	4	+

TABLE II – *continued*

	RPP		*RPE*	
Chemical	*Experimental*	*Prediction*	*Experimental*	*Prediction*
Gibberellic acid	-	+	4	-
2,3,4,5,6-Pentachlorobiphenyl	-	+	4.4	+
2-Monochlorobiphenyl	-	-	4.4	+
2,3,4,4′-Tetrachlorobiphenyl	-	+	4.7	+
2′,3′,4′,5,'5-Pentachloro-2-hydroxybiphenyl	-	+	4.8	+
4-Ethylphenol	-	+	5	+
Chlordane	-	+	5	+
3,5-Dichloro-2-hydroxybiphenyl	-	+	5.4	+
2,3,6-Trichlorobiphenyl	-	-	5.8	+
Heptachlor	-	+	8	+
4-Propylphenol	-	+	17	+
6-Bromonaphthol-2	-	+	38	+
t-Butylhydroxyanisol	0.00006	+	30	+
2′,5′-Dichloro-2-hydroxybiphenyl	0.0001	-	13	+
2′,3′,4′,5′-Tetrachloro-3-hydroxybiphenyl	0.0001	+	35.3	+
2,3,4,5-Tetrachlorobiphenyl	0.0001	-	39.2	+
1-Hydroxychlordene	0.0001	+	40	+
Toxaphene	0.0001	-	51.9	-
Dieldrin	0.0001	-	54.89	+
Methoxychlor	0.0001	+	57	+
2,2′,3,3′,6,6′-Hexachlorobiphenyl	0.0001	-	61.6	+
2,2′,4,5-Tetrachlorobiphenyl	0.0001	+	61.6	+
2′,5′-Dichloro-3-hydroxybiphenyl	0.0001	+	69.9	+
p,p′-DDT	0.0001	+	71	+
2,4,4′,6-Tetrachlorobiphenyl	0.0001	-	75.7	+
2,3,4-Trichlorobiphenyl	0.0001	-	77	+
Endosulfan	0.0001	-	81.25	-
Kepone	0.0001	-	84	-
o,p′-DDD	0.0001	-	84	+
o,p′-DDT	0.0001	+	86.14	+
4-*tert*-Butylphenol	0.0003	+	71	+
4-*sec*-Butylphenol	0.0003	+	76	+
Bisphenol A	0.0003	+	82	+
4,4′-Dihydroxyiphenyl	0.0003	+	84	+
4-Hydroxybiphenyl	0.0003	+	87	+
Butylbenzylphthalate	0.0003	-	90	-
4-*iso*-Pentylphenol	0.0003	-	93	-
4-*tert*-Pentylphenol	0.0003	+	105	+
2,2′,5-Trichloro-4-hydroxybiphenyl	0.001	+	37.8	+
2′,5′-Dichloro-4-hydroxybiphenyl	0.001	+	71.2	+
Tamoxifen	0.001	+	75*	+
2′,3′,4′,5′-Tetrachloro-4-hydroxybiphenyl	0.001	+	92	+
Coumestrol	0.001	+	93	-
Bisphenol A dimethacrylate	0.003	-	84	+
4-Nonylphenol	0.003	+	100	+
2′,4′,6′-Trichloro4-hydroxybiphenyl	0.01	+	99.8	+
4-Octylphenol	0.03	+	100	+
5-Octylphenol	0.03	+	100	+
Pseudo diethylstilbestrol	0.1	+	100	+
16-Hydroxyestrone	0.1	+	-	+
Zearalenone	1	+	88	-
Zearalenol	1	+	93	-
Equilenin	1	+	100	+
Estrone	1	+	100	+
Allenolic acid	1	-	105	-
Estriol	10	+	100	+
Indenestrol	10	+	100	+
17β-estradiol	100	+	100	+
Ethynylestradiol	100	+	100	+
11β-chloromethylestradiol	1000	+	110	+

TABLE II – *continued*

	RPP		*RPE*	
Chemical	*Experimental*	*Prediction*	*Experimental*	*Prediction*
Moxestrole	1000	+	110	+
Diethylstilbestrol	1000	+	112	+

Note:
* RPE potency value estimated from compounds with similar RPP values.

to oxidative mutagens [39] of the type that are derived from estrogens [40–43]. Interestingly, there is antagonism between estrogenicity and the induction of micronuclei as indicated by the significantly less than expected overlap (Table V, Analysis 6). This could reflect that there are two mechanisms to induce micronuclei: genotoxic vs. non-genotoxic (e.g. via inhibition of tubulin polymerization). However, analysis of estrogens and carcinogens

TABLE III MCASE biophores associated with estrogenic activity measured by RPE in the ESCREEN assay

Biophore	*Total*	*Inactive*	*Active*
1a. cH =cH –c =cH –	70	13	57
1b. c ⟨ =cH –c =c ⟩ –	2	0	2
1c. cH = c –c = c–	24	0	24
1d. cH = cH –c =c –	26	0	26
2. Cl –c =c –c =	24	0	24
3. Cl –C –C =	7	0	7
4. OH –CH –	7	0	7
5. CO –C –	5	0	5

Each biophore is accompanied by the number of compounds contributing to it, the number of active and inactive compounds, and their average activity.
Notes:
Biophore interpretation:
c: aromatic carbon.
⟨⟩: attachment of electron withdrawing or electron donating group.
ˆ: epoxide.
(#-atom): biophore branch at atom # with substituent.
See Fig. 1 for illustration of biophores.

TABLE IV MCASE biophores associated with estrogenic activity measured by RPP in the ESCREEN assay

Biophore		*Total*	*Inactive*	*Active*
1a. cH =cH –c ⟨ =cH –		41	9	32
1b. c =cH –c ⟨ =c –		1	0	1
1c. c ⟨=cH –c =c ⟩–		2	0	2
2a. cH =c –c =c –c =cH –	⟨2-Cl ⟩	4	0	4
2b. cH =c –cH =c –c =c –	⟨2-Cl ⟩	2	0	2
3a. cH =cH –c =cH –cH =c –CH –		6	0	6
3b. CH2 –c =cH –cH =cH –cH =cH –		1	0	1
5. Cl –c =c –c =c –c =cH –cH =cH –cH = ⟨5-cH= ⟩		2	0	2
6. C –C –C –C – ⟨2-Cl ⟩		3	1	2
7. OH –CH –CH –		2	0	2
8. Ô –CH –		1	0	1
9. O –SO –O –CH2–CH –C –C –	⟨6-Cl ⟩	1	0	1

Each biophore is accompanied by the number of compounds contributing to it, the number of active and inactive compounds, and their average activity.
Notes: see Table III.
See Fig. 2 for illustration of biophores.

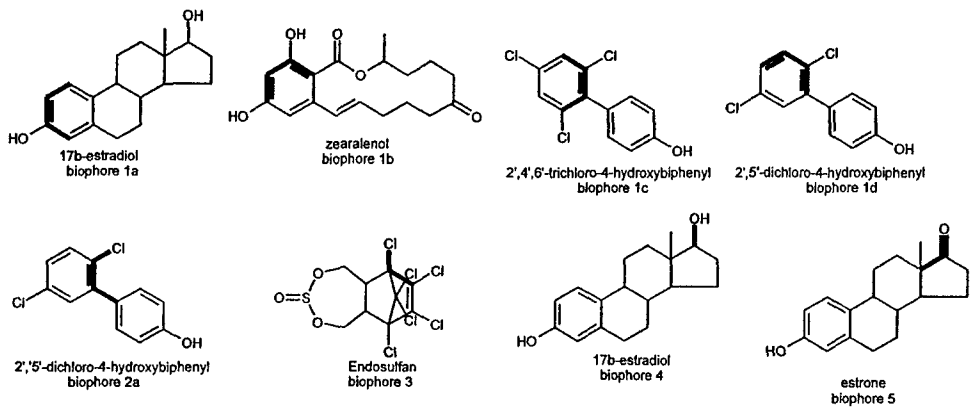


FIGURE 1 Illustration of MCASE biophores associated with estrogenic activity measured by RPE in the ESCREEN assay.

indicate that they may significantly share a common underlying mechanism (Table V, Analyses 7–10). Only the CPDB rat model did not significantly overlap with estrogenicity (Table V, Analysis 8). Analyses 11 and 12 of Table V show a significant overlap between estrogenicity and developmental toxicity in both humans and hamsters. Overall, these findings provide credibility to the mechanistic basis of the ESCREEN models.

CONCLUSIONS

The present analysis of estrogenicity with the MCASE program clearly indicates the utility of the program in assessing unknown compounds for estrogenicity. Given the complex structural nature of estrogenic compounds, it is imperative that any computational method

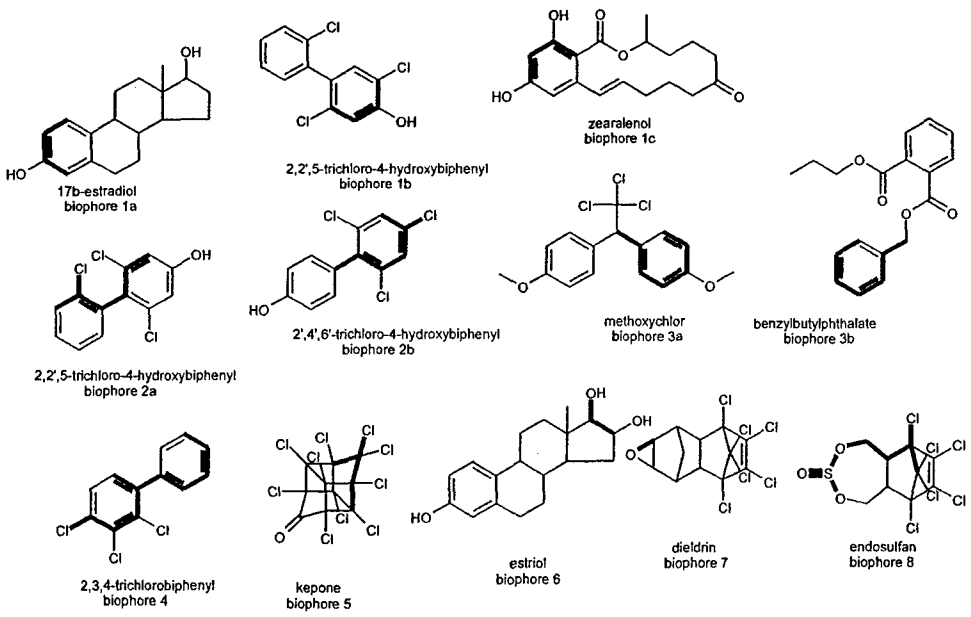


FIGURE 2 MCASE biophores associated with estrogenic activity measured by RPP in the ESCREEN assay.

TABLE V Mechanistic relationships of the ESCREEN RPE assay to other toxicological endpoints including genotoxicity, developmental effects and carcinogenesis

Analysis (References)	*Observed*	*Expected*	*p-value*	*Δ**	*100Δ/Expected*
1. ESCREEN relative proliferative potency	776	236	< 0.0001	540	228.8
2. Salmonella mutagenicity [44,45]	470	461	0.763	9	1.9
3. Unscheduled DNA synthesis [46]	485	445	0.179	40	9.0
4. Chromosomal aberrations [47]	364	400	0.184	− 36	− 9
5. SOS chromotest [48,49]	338	273	< 0.0001	65	23.8
6. Induction of micronuclei [50]	16	125	< 0.0001	− 109	− 87.2
7. CPDB mouse [37]	561	466	0.002	95	20.4
8. CPDB rat [38]	531	490	0.188	41	8.4
9. NTP mouse [47]	843	555	< 0.0001	288	51.9
10. NTP rat [47]	407	271	< 0.0001	136	50.2
11. Hamster developmental toxicity [51]	491	416	0.011	75	18.0
12. Human developmental toxicity [52]	338	274	0.009	64	23.4

Notes:
Observed: Number of compounds simultaneously identified to be estrogens using the RPE model and the row-listed endpoint.
Expected: The product of the individual prevalences of compounds identified to be estrogens using the RPE model and the row-listed endpoint.
p-value: Difference of two means test.
Δ: Difference of observed from expected.
100Δ/Expected: Percent difference from expected.

applied to their analysis is capable of coping with noncongeneric datasets. As evidenced by MCASE's predictive performance, it seems likely that this program has the potential to be a useful tool for screening and prioritizing environmental agents for subsequent testing. Moreover, we are applying this method in the development of models depicting relative binding ability to the estrogen receptor and for uterotropic and antiuterotropic activity. We speculate that although the program could be used as a stand-alone entity for screening potentially endocrine active compounds, it seems more likely and prudent that it could contribute as part of a battery of computational tools aimed at prioritizing compounds.

Acknowledgements

We gratefully acknowledge support for this work from the Congressionally Directed Medical Research Program for Breast Cancer Idea Award DAMD17-01-0376.

References

[1] Adlercreutz, H. (1993) "Phytoestrogens: Epidemiology and a possible role in cancer protection", *Environ. Health Perspect.* **103**(Suppl 7), 103–112.
[2] Clarkson, T.B., Anthony, M.S., Williams, J.K., Honore, E.K. and Cline, J.M. (1998) "The potential of soybean phytoestrogens for postmenopausal hormone replacement therapy", *Proc. Soc. Exp. Biol. Med.* **217**, 365–368.
[3] Arjmandi, B.H. (2001) "The role of phytoestrogens in the prevention and treatment of osteoporosis in ovarian hormone deficiency", *J. Am. Coll. Nutr.* **20**, 398s–402s.
[4] Marselos, M. and Tomatiz, L. (1992) "Diethylstilbestrol: I, Pharmacology, toxicology and carcinogenicity in humans", *Eur. J. Cancer* **28A**, 1182–1189.
[5] Marselos, M. and Tomatiz, L. (1993) "Diethylstilbestrol: II, Pharmacology, toxicology and carcinogenicity in experimental animals", *Eur. J. Cancer* **29A**, 149–155.
[6] IARC (1979) Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans, Sex Hormones (II) (International Agency for Research on Cancer, Lyon) Vol. **21**.
[7] vom Saal, F.S., Montano, M.M. and Wang, M.H. (1992) "Sexual differentiation in mammals", In: Colborn, T. and Clement, C., eds, Chemically-Induced Alterations in Sexual Development: The Wildlife/Human Connection (Princeton Scientific Publishing, Princeton, NJ), pp 17–84.
[8] Gray, J.L.E. (1992) "Chemical-induced alterations of sexual differentiation: a review of effects in humans and rodents", In: Colborn, T. and Clement, C., eds, Chemically-Induced Alterations in Sexual Development: The Wildlife/Human Connection (Princeton Scientific Publishing, Princeton, NJ), pp 203–230.

[9] Blair, B.B. (1992) "Immunologic studies of women exposed *in utero* to diethylstilbestrol", In: Colborn, T. and Clement, C., eds, Chemically-Induced Alterations in Sexual Development: The Wildlife/Human Connection (Princeton Scientific Publishing, Princeton, NJ), pp 289–294.
[10] Katzenellenbogen, J.A. (1995) "The structural pervasiveness of estrogenic activity", *Environ. Health Perspect. Suppl.* **103**(Suppl 7), 99–101.
[11] EPA (1998) "Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC) Final Report", [www.http://www.epa.gov/oscpmont/oscpendo/history/finalrpt.htm].
[12] EPA (2002) "Priority-Setting in the Endocrine Disruptor Screening Program (EDSP)-Background. Washington, DC: Environmental Protection Agency", [www.http://www.epa.gov/oscpmont/oscpendo/prioritysetting/background.htm].
[13] Timm, G (2002) "EPA update on the validation and standardization, In: TestSmart Endocrine Disruptors", [www.http://caat.jhsph.edu/programs/workshops/testsmart/endo02-proc.htm].
[14] Cronin, M.T.D., Jaworska, J.S., Walker, J.D., Comber, M.H.I., Watts, C.D. and Worth, A.P. (2003) "Use of quantitative structure–activity relationships in international decision-making frameworks to predict health effects of chemical substances", *Environ. Health Perspect.* **111**, 1376–1390.
[15] Cronin, M.T.D., Walker, J.D., Jaworska, J.S., Comber, M.H.I., Watts, C.D. and Worth, A.P. (2003) "Use of quantitative structure–activity relationships in international decision-making frameworks to predict ecological effects and environmental fate of chemical substances", *Environ. Health Perspect.* **111**, 1376–1390.
[16] Waller, C.L., Oprea, T.I., Chae, K., Park, H.-K., Korach, K.S., Laws, S.C., Wiese, T.E., Kelce, W.R. and Gray, J.L.E. (1996) "Ligand-based identification of environmental estrogens", *Chem. Res. Toxicol.* **9**, 1240–1248.
[17] Tong, W., Perkins, R., Strelitz, R., Collantes, E.R., Keenan, S., Welsh, W.J., Branham, W.S. and Sheehan, D.M. (1997) "Quantitative structure–activity relationships (QSARs) for estrogen binding to the estrogen receptor: predictions across species", *Environ. Health Perspect.* **105**, 1116–1124.
[18] Tong, W., Perkins, R., Xing, L., Welsh, W.J. and Sheehan, D.M. (1997) "QSAR models for binding of estrogenic compounds to estrogen receptor alpha and beta subtypes", *Endocrinology* **138**, 4022–4025.
[19] Tong, W., Lowis, D.R., Perkins, R., Chen, Y., Welsh, W.J., Goddette, D.W., Heritage, T.W. and Sheehan, D.M. (1998) "Evaluation of quantitative structure–activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor", *J. Chem. Inf. Comput. Sci.* **38**, 669–677.
[20] Blair, R.M., Fang, H., Branham, W.S., Hass, B.S., Dial, S.L., Moland, C.L., Tong, W., Shi, L., Perkins, R. and Sheehan, D.M. (2000) "The estrogen receptor relative binding affinities of 188 natural and xenochemicals: Structural diversity of ligands", *Toxicol. Sci.* **54**, 138–153.
[21] Shi, L.M., Fang, H., Tong, W., Wu, J., Perkins, R., Blair, R.M., Branham, W.S., Dial, S.L., Moland, C.L. and Sheehan, D.M. (2001) "QSAR models using a large diverse set of estrogens", *J. Chem. Inf. Comput. Sci.* **41**, 186–195.
[22] Hong, H., Tong, W., Fang, H., Shi, L., Xie, W., Wu, J., Perkins, R., Walker, J.D., Branham, W. and Sheehan, D.M. (2002) "Prediction of estrogen receptor binding for 58,000 chemicals using an integrated system of a tree-based model with structural alerts", *Environ. Health Perspect.* **110**, 29–36.
[23] Soto, A.M., Lin, T.-M., Justicia, H., Silvia, R.M. and Sonnenschein, C. (1992) "An "in culture" bioassay to assess the estrogenicity of xenobiotics (E-SCREEN)", In: Colborn, T. and Clement, C., eds, Chemically-Induced Alterations in Sexual Development: The Wildlife/Human Connection (Princeton Scientific Publishing, Princeton, NJ), pp 295–309.
[24] Soto, A.M., Sonnenschein, C., Chung, K.L., Fernandez, M.F., Olea, N. and Serrano, F.O. (1995) "The E-SCREEN assay as a tool to identify estrogens: An update on estrogenic environmental pollutants", *Environ. Health Perspect.* **103**(Suppl 7), 113–122.
[25] Sonnenschein, C., Soto, A.M., Fernandez, M.F., Olea, N., Olea-Serrano, M.F. and Ruiz-Lopez, M.D. (1995) "Development of a marker of estrogenic exposure in human serum", *Clin. Chem.* **41**, 1888–1895.
[26] Klopman, G. (1984) "Artificial intelligence approach to structure–activity studies. Computer automated structure evaluation of biological activity of organic molecules", *J. Am. Chem. Soc.* **106**, 7315–7321.
[27] Klopman, G. (1992) "MULTICASE 1. A hierarchical computer automated structure evaluation program", *Quant. Struct. Act. Relat.* **11**, 176–184.
[28] Klopman, G. and Rosenkranz, H.S. (1994) "Approaches to SAR in carcinogenesis and mutagenesis. Prediction of carcinogenicity / mutagenicity using MULTI-CASE", *Mutat. Res.* **305**, 33–46.
[29] Cunningham, A.R., Klopman, G. and Rosenkranz, H.S. (1996) "The carcinogenicity of diethylstilbestrol: structural evidence for a non-genotoxic mechanism", *Arch. Toxicol.* **70**, 356–361.
[30] Murrill, W.B., Brown, N.M., Zhang, J.-X., Manzolillo, P.A., Barnes, S. and Lamartiniere, C.A. (1996) "Prepubertal genistein exposure suppresses mammary cancer and enhances gland differentiation in rats", *Carcinogenesis* **17**, 1451–1457.
[31] Macina, O.T., Zhang, Y.P. and Rosenkranz, H.S. (1998) "Improved predictivity of carcinogens: the use of a battery of SAR models", In: Kitchin, K., ed, Testing, Predicting and Integrating Carcinogenicity (Marcel Dekker, New York), pp 227–250.
[32] Chankong, V., Haimes, Y.Y., Rosenkranz, H.S. and Pet-Edwards, J. (1985) "The carcinogenicity prediction and battery selection (CPBS) method: a Bayesian approach", *Mutat. Res.* **153**, 135–166.
[33] Zhang, Y.P., Sussman, N., Klopman, G. and Rosenkranz, H.S. (1997) "Development of methods to ascertain the predictivity and consistency of SAR models: Application to the U.S. National Toxicology Program rodent carcinogenicity bioassays", *Quant. Struct. Act. Relat.* **16**, 290–295.
[34] Pollack, N., Cunningham, A.R., Klopman, G. and Rosenkranz, H.S. (1999) "Chemical diversity approach for evaluating mechanistic relatedness among toxicological phenomena", *SAR QSAR Environ. Res.* **10**, 533–543.

[35] Rosenkranz, H.S. and Cunningham, A.R. (2001) "Prevalence of mutagens in the environment: experimental data vs. simulation", *Mutat. Res.* **484**, 49–51.
[36] Shi, L.M., Fan, Y., Myers, T.G., O'Connor, P.M., Paull, K.D., Friend, S.H. and Weinstein, J.N. (1998) "Mining the NCI anticancer drug discovery database: genetic function approximation for the QSAR of anticancer ellipticine analogues", *J. Chem. Inf. Comput. Sci.* **38**, 189–199.
[37] Cunningham, A.R., Rosenkranz, H.S., Zhang, Y.P. and Klopman, G. (1998) "Identification of "genotoxic" and "non-genotoxic" alerts for cancer in mice: the carcinogenic potency database", *Mutat. Res.* **398**, 1–17.
[38] Cunningham, A.R., Rosenkranz, H.S. and Klopman, G. (1998) "Identification of structural features and associated mechanisms of action for carcinogens in rats", *Mutat. Res.* **405**, 9–28.
[39] Rosenkranz, H.R. (2002) "A paradigm for determing the relevance of short-term assays: Application to oxidative mutagenesis", *Mutat. Res.* **508**, 21–27.
[40] Sipe, H.J., Jr., Jordan, S.J., Hanna, P.M. and Mason, R. (1994) "The metabolism of 17b-estradiol by lactoperoxidase: a possible source of oxidative stress in breast cancer", *Carcinogenesis* **15**, 2637–2643.
[41] Liehr, J.G., DaGua, B.B. and Ballatore, A.M. (1985) "Reactivity of 4′,4-diethystilbestrol quinone, a metabolic intermediate of diethylstilbestrol", *Carcinogenesis* **6**, 829–836.
[42] Roy, D., Floyd, R.A. and Liehr, J.G. (1991) "Elevated 8-hydroxyguanosine levels in DNA of diethylstilbestrol-treated Syrian Hamsters: Covalet DNA damage by free radicals generated by redox cycling of diethylstilbestrol", *Cancer Res.* **51**, 3882–3885.
[43] Liehr, J.G. (1990) "Genotoxic effects of estrogens", *Mutat. Res.* **238**, 269–276.
[44] Zeiger, E., Ashby, J., Bakale, G., Enslein, K., Klopman, G. and Rosenkranz, H.S. (1996) "Prediction of Salmonella mutagenicity", *Mutagenesis* **11**, 471–484.
[45] Lui, M., Sussman, N., Klopman, G. and Rosenkranz, H.S. (1996) "Estimation of the optimal database size for structure-activity analyses: the Salmonella mutagenicity database", *Mutat. Res.* **358**, 63–72.
[46] Rosenkranz, H.S., Zhang, Y.P. and Klopman, G. (1994) "Evidence that cell toxicity may contribute to the genotoxic response", *Regul. Toxicol. Pharmacol.* **19**, 176–182.
[47] Ennever, F.K., Rosenkranz, H.S., Lave, L.B. and Omenn, G.S. (1990) "Value-of-information analysis of testing strategies: estimating the effect of uncertainty about the proportion of chemicals that are true human carcinogens", In: Mendelsohn, M.L. and Albertini, R.J., eds, Mutation and the Environment, Part D: Carcinogenesis (Wiley-Liss, Hoboken, NJ), pp 23–48.
[48] Mersch-Sundermann, V., Klopman, G. and Rosenkranz, H.S. (1996) "Chemical structure and genotoxicity: studies of the SOS Chromotest", *Mutat. Res.* **340**, 81–91.
[49] Mersch-Sundermann, V., Schneider, U., Klopman, G. and Rosenkranz, H.S. (1994) "SOS-Induction in *E. coli* and *Salmonella* mutagenicity: a comparison using 330 compounds", *Mutagenesis* **9**, 205–224.
[50] Yang, W.-L., Klopman, G. and Rosenkranz, H.S. (1992) "Structural basis of the *in vivo* induction of micronuclei", *Mutat. Res.* **272**, 111–124.
[51] Gómez, J., Macina, O.T., Mattison, D.R., Zhang, Y.P., Klopman, G. and Rosenkranz, H.S. (1999) "Structural determinants of developmental toxicity in hamsters", *Teratology* **60**, 190–205.
[52] Ghanooni, M., Mattison, D.R., Zhang, Y.P., Macina, O.T., Rosenkranz, H.S. and Klopman, G. (1997) "Structural determinants associated with risk of human developmental toxicity", *Am. J. Obstet. Gynecol.* **176**, 799–806.